Análisis de Estructura Temporal de Datos Musicales para Clasificación

María Victoria López García
Ingeniería Técnica de Sistemas de Telecomunicación
Universidad Carlos III – Leganés - Madrid

Tabla de contenido

Índice de figuras

Resumen

La Recuperación de Información Musical constituye un campo de investigación muy activo que se nutre de muy diversas disciplinas, como Musicología, Acústica, Psicología o Aprendizaje Máquina. En el ámbito académico el resultado es la edición anual de numerosas publicaciones. La industria musical también ha sufrido una gran transformación en las últimas décadas debido a las nuevas reglas que ha impuesto la distribución digital. Actualmente la mayoría de los usuarios puede acceder a grandes colecciones de canciones y escucharla en los dispositivos móviles existentes (reproductores MP3, iPods, smartphones, etc.). En este contexto es imprescindible desarrollar sistemas que ayuden a los usuarios a organizar estas bases de datos musicales o a acceder fácilmente a canciones acordes a sus preferencias. Los sistemas de recomendación han ganado una gran popularidad y son ya parte habitual en sitios de compra por Internet. En los sitios de descarga musical son capaces de generar listas de reproducción basándose en las descargas previas del usuario.

Estos sistemas de recomendación automática normalmente están fundamentados en algoritmos de aprendizaje máquina que son capaces de extraer información relevante de las canciones tras una fase de entrenamiento; para ello se consideran varias tareas cuyo objetivo es extraer información de bajo y alto nivel de la señal de audio. Este Proyecto se ha centrado en la extracción de descriptores de alto nivel y en estudiar un esquema utilizado previamente en clasificación de género musical, reconocimiento de portadas de álbumes, reconocimiento de instrumentos, etc. La mayor dificultad a la que se enfrentan estos esquemas es la de cubrir el escalón existente entre los descriptores de bajo nivel, que se pueden extraer de los archivos de audio (usualmente bajo la forma de Coeficientes Cepstrales de Frecuencia en escala Mel, MFCC) y la información semántica que pretenden recuperar (por ejemplo, el género de la canción).

Los sistemas que se van a estudiar en este Proyecto se basan en la concatenación de los siguientes pasos: extracción de descriptores de bajo nivel, integración temporal de características y clasificación automática. La primera etapa, la extracción de descriptores de bajo nivel, está implementada como extracción de coeficientes MFCC, si bien también se podrían considerar otros descriptores. A pesar de que originalmente fueron propuestos para tareas de reconocimiento de voz, los coeficientes MFCC han ganado su propio espacio en la Recuperación de Información Musical debido a su excelente desempeño en sistemas basados en representación espectral. El objetivo de la fase de integración temporal es concentrar la información de los MFCC de varias ventanas consecutivas de corta duración (la escala habitual es de 20-40 ms.) en un único vector de características en el que está representada la información más relevante a una escala de tiempos mayor. Este punto es de gran importancia, pues la información de alto nivel requerida sólo se puede detectar en esta nueva escala temporal, por tanto la simple concatenación de coeficientes MFCC de ventanas adyacentes implicaría serios inconvenientes prácticos en la siguiente fase. Finalmente, la etapa de aprendizaje máquina utiliza los datos de la etapa anterior como conjunto de entrenamiento para crear un modelo matemático cuyo propósito es predecir la información de interés de los vectores de características. En este Proyecto se han considerado como herramienta para superar esta fase los clasificadores no lineales basados en métodos de núcleos.

El trabajo desarrollado en este Proyecto se concentra en la integración temporal de características y en estudiar la influencia de la duración de la ventana temporal en el proceso de clasificación musical. La hipótesis inicial es que la duración de la ventana óptima para la fase de integración depende de la tarea particular para la que el sistema está diseñado. Por esta razón se consideran dos bases de datos de canciones diferentes, una está diseñada para la clasificación de género y la otra para clasificar en función de los artistas que interpretan las canciones. Se estudiará cómo la precisión de la clasificación cambia en función de la duración de la ventana utilizada para la integración de características, a la vez que intenta arrojar alguna luz sobre la importancia de este parámetro para el correcto desempeño del sistema. La conclusión principal de este Proyecto es que una selección incorrecta del tamaño de la ventana temporal puede conducir a una funcionalidad deficiente. Por tanto, parece crucial validar correctamente su valor durante el diseño del sistema de clasificación.

Abstract

Music Information Retrieval (MIR) constitutes a very active research field, with many papers published every year with contributions coming from several disciplines, such as musicology, acoustics, psychology, or machine learning. The music industry has also transformed during the last decades as a consequence of digital distribution. Nowadays, most users have access to huge collections of songs, and can even carry them on portable devices (MP3 players, ipods, smartphones, etc). In this context, it becomes crucial to develop systems that can help the users to organize these music databases, or to access songs according to each personal user's preferences. Recommender systems have gained popularity and are now common in e-commerce sites and they are even implemented in playing software in the form of automatic reproducing list generators.

These automatic recommendation systems are normally based on machine learning algorithms that are trained to extract relevant information from the music. In this sense, several tasks have been considered to extract low-level and high-level information from the audio waveform. In this work, we focus on the extraction of high-level descriptors and study a classification scheme that has been used for this purpose in tasks such as genre classification, album cover detection, instrument detection, etc. The main difficulty that face these schemes is to fill the gap that exists between the low-level descriptors that can be extracted from the audio file (normally in the form of Mel Frequency Cepstral Coefficients, MFCCs) and the semantic information they want to retrieve (e.g., the genre).

The systems we study in this work consist of the concatenation of the following steps: low-level descriptor extraction, temporal feature integration, and automatic classification. The first stage, low-level descriptor extraction, is implemented in the form of MFCC extraction, although other descriptors could be considered as well. In spite of being originally proposed for speech recognition tasks, MFCCs have gained popularity in MIR because of the reported good performance of systems that are based on this spectral representation. The goal of the time integration phase is to concentrate the information of MFCCs from several consecutive short-time windows (normally in the scale of 20-40 ms), and produce a new representation in the form of a unique feature vector capturing the most relevant information at a larger time-scale. This is important because the high-level information we are looking for normally can only be detected in this larger time-scale, and the simple concatenation of MFCCs from adjacent windows would imply practical inconveniences from the point of view of the machine learning step. Finally, the machine learning module uses a training dataset of songs and creates a mathematical model that can be used to predict the information of interest from the time-integrated feature vector. In this work, non-linear classifiers based on kernel methods will be considered as the classification technology.

This project work concentrates on the time integration of features, and studies the influence of the window length on the performance of the overall filter. It is expected that the length of the optimal window for the time integration phase depends on the particular task the system is designed for. For this reason, we consider two different song databases, one of them is designed for genre classification tasks, while the goal of the second one is to classify songs according to the artist. We will study how the classification accuracy changes with the length of the window used for time-feature integration, and shed some light into the importance of this parameter for the success of the overall filter. The main conclusion of our work is that a bad selection of this parameter can lead to very suboptimal performance; consequently, it seems crucial to validate its value during the design of the system.

CAPÍTULO 1. Panorámica sobre la clasificación musical

1.1. Introducción

La música existe desde la noche de los tiempos. Todavía no ha sido descubierta ninguna cultura en la que no se haya manifestado de forma más o menos sofisticada, por lo que se podría concluir que este arte, junto con otras expresiones de toda índole, no sólo complementan a la cultura, sino que la determinan.

A lo largo de la historia la música ha desempeñado numerosas funciones: El preámbulo del rito de la caza para el hombre prehistórico, los coros en la tragedias griegas, el canto de los trovadores animando la vida de las aldeas o su consumo masivo gracias a la aparición de Internet… se podría continuar hasta construir una lista tan extensa en su longitud como compleja en la enumeración de desempeños. La relación del hombre con la música ha variado con el paso de los siglos, pero siempre manteniendo su carácter de exteriorización del sentimiento humano.

Actualmente, como ya se ha mencionado, la música se vive como un bien de consumo masivo, aunque simultáneamente constituye un elemento que personaliza y distingue a sus oyentes; sólo desde esta perspectiva podría explicarse el auge de las tecnologías de recuperación de datos musicales - MIR, Music Information Retrieval –, tanto desde el punto de vista de la investigación como del de la aplicación.

Este proyecto dará una visión técnica de la clasificación de audio en función del género musical de la canción y el artista que la interpreta, centrándose en analizar en qué horizonte temporal puede extraerse de forma más precisa la información sobre cada una de las dos categorías (o como se expondrá más adelante, características de alto nivel); para ello, se analizarán dos bases de datos cuyo contenido son archivos de audio etiquetados de acuerdo al género o artista, respectivamente. Con cada una de ellas se realizará una serie de experimentos empleando diferentes técnicas de extracción de características y de integración temporal y variando los parámetros que determinan a cada una, entre ellos y de forma preponderante el tamaño de la ventana temporal. Tras la finalización de la etapa de experimentación se evaluará la tasa de reconocimiento de la máquina y se podrá concluir con qué tamaño de ventana temporal se extrae de forma más exacta determinada característica; resumiendo, el objetivo será conocer de forma más precisa la relevancia y criticidad de distintos parámetros y entender mejor en qué escala temporal radica la información de interés.

1.1.1. Organización del Capítulo

En este capítulo se presentan algunas generalidades acerca del proceso de clasificación del contenido musical, cuyo diagrama de bloques se ilustra en la Figura 1. Esencialmente, la estructura de este capítulo introductorio será paralela a los procedimientos enumerados en dicho diagrama. En capítulos posteriores del proyecto se analizarán con más detalle cada una de las etapas que componen dicho proceso.

Figura 1. Proceso de clasificación general

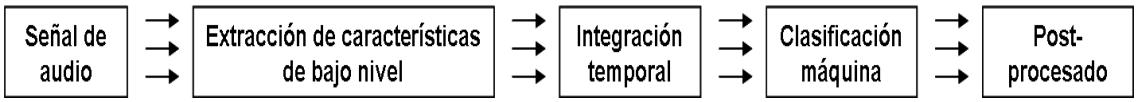


Diagrama de bloques del proceso de clasificación de contenido musical. El objetivo de este proceso será obtener una característica de alto nivel, intuitiva para el ser humano. Para ello se extraen las características de bajo nivel de la señal de audio utilizando distintos horizontes temporales; posteriormente estas características se reúnen en un vector de horizonte temporal mayor, hecho que constituye la fase de integración temporal. La etapa de clasificación máquina, en su más amplio sentido, supone el entrenamiento y validación de una máquina para que sea capaz de deducir las características de alto nivel a partir del vector de entrada. Por último, el post-procesado implica optimizar o mejorar la salida de la clasificación máquina.

En este punto es importante remarcar que la extracción de características se ocupa de obtener las de bajo nivel, esto es, las descripciones de contenido que tienen una implicación física y pueden ser obtenidas mediante algoritmos, como el timbre, el tono, estimaciones de los armónicos, etc. Éstas deben contraponerse a las de alto nivel, que requieren un nivel de abstracción elevado y en cuya elaboración se deja ver la huella humana, como por ejemplo el género de una canción [Detyniecki et al., 2005]. El objetivo del proceso será inferir las características de alto nivel a partir de las de bajo nivel.

Las características de bajo nivel son eminentemente físicas y no intuitivas para el ser humano; sin embargo, las de alto nivel constituyen conceptos con una semántica ya elaborada que las personas manejan con mayor soltura, y que por otra parte poseen cierto grado de subjetividad al estar definidos bajo el prisma personal. La distancia conceptual entre ambos tipos de características se denomina "salto semántico" ("semantic gap") y se erige como materia de estudio en la Lingüística, la Lógica y la Informática, ciencias en las que los lenguajes formales adquieren una importancia innegable.

Con carácter previo, en este capítulo se ofrecerá una panorámica de sistemas de organización existentes, con el fin de contextualizar el ámbito de trabajo. Tras esto se esbozarán los conceptos e ideas que sustentan el desarrollo de este Proyecto, siguiendo el orden del diagrama expuesto anteriormente. Por último, y para facilitar el seguimiento de esta memoria, se expondrá la distribución de los siguientes capítulos.

1.2. Sistemas comerciales de organización de música

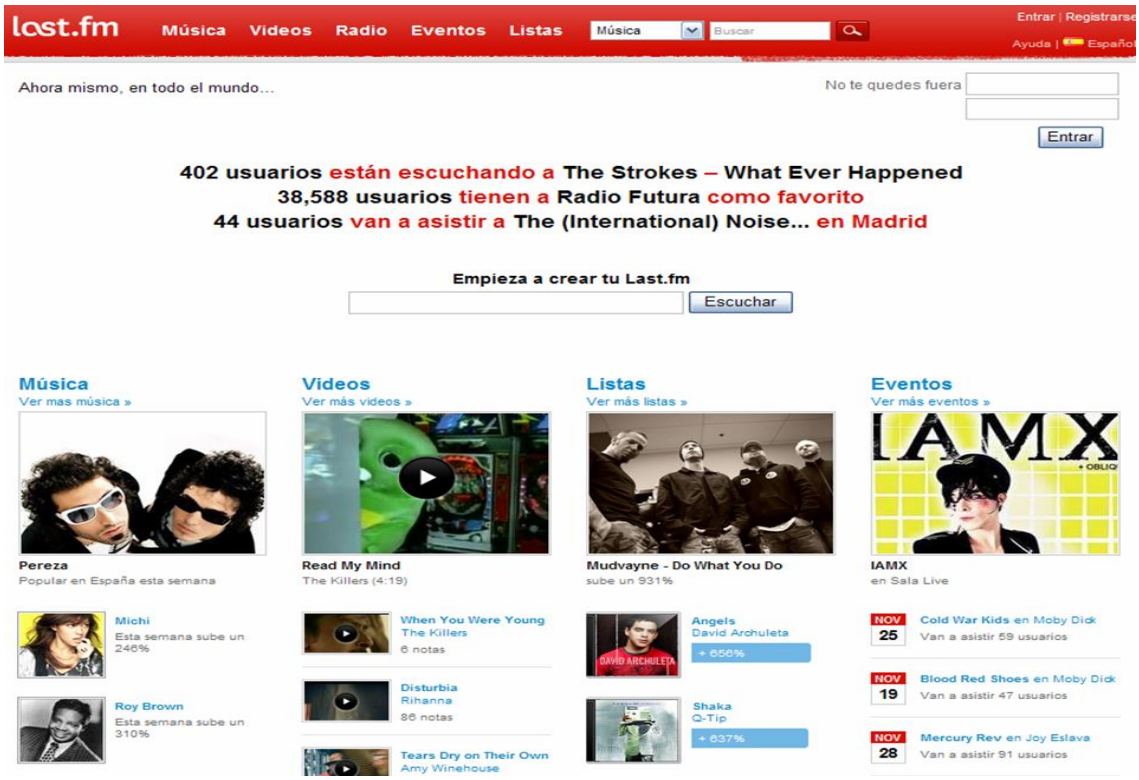
La enorme cantidad de bases de datos musicales disponibles en la red, así como los dispositivos portátiles de reproducción existentes, capaces de almacenar hasta centenas de gigabytes, hacen imprescindible la existencia de sistemas de organización musical. En la red pueden encontrarse ejemplos de bases de datos musicales que ofrecen diversas funcionalidades:

- Last.fm (www.last.fm). Sistema de recomendación de canciones que construye perfiles y estadísticas sobre gustos musicales. El sistema de recomendación de canciones se denomina 'Audioscrobbler'. Asimismo, tiene funcionalidades de red social y radio via Internet.

  Last.fm es un sistema de recomendación de canciones que funciona en muchos de los reproductores de música existentes en varios sistemas operativos, como Windows o Mac. Un usuario de Last.fm puede construir un perfil musical usando dos métodos: bien escuchando su colección musical personal en una aplicación de música con un plug-in, o bien utilizando su servicio de radio a través de Internet, normalmente a través del reproductor que proporciona la propia página, aunque es sus últimas versiones también ofrece la posibilidad de utilizar uno externo. Last.fm crea perfiles a partir de las selecciones musicales de diversos usuarios, suponiendo que si dos o más usuarios comparten una elección es probable que tengan más preferencias en común. Asimismo dispone de temas etiquetados que permiten al usuario crear bases de datos en función del estilo al tiempo que le permite subir sus propias canciones, permitiendo que el resto de los usuarios disfruten de ellas a través de la descarga gratuita.

  Las recomendaciones son calculadas usando un algoritmo colaborativo de filtrado, así los usuarios pueden explorar una lista de artistas no listados en su propio perfil pero que sí que aparecen en otros usuarios con gustos similares. Last.fm también permite la recomendación directa a otros usuarios de discos incluidos en la base de datos de Last.fm. El stock musical de Last.fm contiene más de 100.000 canciones.

Figura 2. Página de entrada de Last.fm



Página de entrada de last.fm (Fuente: www.last.fm)

- Pandora (www.pandora.com). Sistema de recomendación de canciones y de emisoras de radio personalizadas a través de Internet.

  Pandora es una base de datos musical con más de 10.000 índices en la que las canciones se caracterizan en función de elementos como la melodía, el ritmo, la instrumentación, etc. La extracción de estas características fue un trabajo artesanal indescriptiblemente arduo, pues se llevó a cabo de forma completamente manual. Basta con teclear el nombre de una canción o artista en Pandora y en la base de datos se buscarán los índices que se asimilen. Pandora también ofrece la posibilidad de ayudar al programa a aprender una preferencia musical determinada mediante la escucha de selecciones musicales realizadas por un usuario, así como de crear hasta 100 canales, cada uno con un estilo musical diferente. Las elecciones realizadas por Pandora pueden gustar o no al usuario, pero es indudable que constituye una manera de ampliar horizontes musicales y de divertirse, visto el anecdotario que circula por Internet narrando las variopintas relaciones entre las entradas y las salidas. El funcionamiento de las emisoras de radio personalizadas es el siguiente: El usuario sugiere el nombre de una canción la página genera una estación de radio con una lista de canciones cuyos patrones son similares a la canción introducida. Tras esta primera aproximación, el usuario puede ir personalizando dicha lista.

  Pandora es un servicio gratuito que se financia con la publicidad, sin embargo es posible adquirir una versión de pago sin ella. A diferencia de Last.fm, Pandora es un servicio de streaming, pero ni de descarga ni de intercambio.

Figura 3. Página de entrada de Pandora

Página de entrada de Pandora (Fuente: http://beyondtheonewayweb.files.wordpress.com/2008/01/pandora.jpg)

- iTunes. (www.apple.com/es/itunes/*)*. Reproducción, organización y venta de archivos de audio.

  iTunes es una aplicación gratuita para Mac y PC capaz de reproducir música digital y vídeos. Con ella se pueden crear colecciones personalizadas de música y vídeos y navegar por ellas, ofreciendo una gran gama de funcionalidades: mezclar canciones, escuchar canciones de otros ordenadores de la red, compartir listas de reproducción, obtener recomendaciones musicales, grabar CDs, sincronizar contenidos con iPod o iPhone, etc. Otra funcionalidad añadida es la tienda de música online, que dispone de más de 8 millones de ítems. Con otros dispositivos se amplía el abanico de posibilidades, por ejemplo pueden descargarse aplicaciones para iPhone o iPod Touch, así como sincronizar el contenido del iPod, iPhone y Apple TV.

Figura 4. Ejemplo de aplicación de iTunes

Ejemplo de aplicación de iTunes. Fuente: http://www.apple.com/es/itunes/whatis/

- Napster (free.napster.com). Servicio de distribución de archivos mp3.

  Pionero en las redes P2P - aunque no enteramente, pues era preciso utilizar servidores centrales tanto para mantener las listas de sistemas conectados como para dar continuidad al flujo de archivos -, Napster consiguió su máximo apogeo en febrero de 2001; sin embargo, las denuncias ante las violaciones de derechos de autor y la posterior pérdida de esta causa ante los juzgados obligaron a Napster a pagar a las empresas discográficas 26 millones de dólares por daños y otros 10 millones de dólares por futuras licencias. En 2008 Napster anunció el lanzamiento de la tienda más grande y más detallada de mp3 del mundo, con 6 millones de canciones. En 2011 Napster se fusionó con Rhapsody y empezó a operar como servicio de pago en diversos países de América y Europa.

- Musicboxes especializadas ateniéndose a criterios geográficos o de género musical (www.musicboxla.com , www.dojo.ie/musicbox).

Una técnica de empleo muy común en estos sistemas puede describirse, en pocas palabras, como un mapa genético de las distintas canciones. Cada una de las variables de las canciones – relacionadas con la melodía, armonía, ritmo, instrumentación, orquestación, arreglos, letras y todo lo relacionado con el canto y la armonía vocal – se asimila a un gen. Estos genes se acaban ensamblando a la manera del genoma, constituyendo así el "material genético" de la canción. Mediante el análisis y extracción de parámetros de canciones de todos los géneros y orígenes es posible sugerir al usuario, por ejemplo, qué otras canciones podrían ser de su agrado basándose en las similitudes entre el "mapa genético" de éstas y el de una canción seleccionada previamente por el usuario; esto es lo que se denomina "motor de recomendación" (*recommendation engine*) y se utiliza, por ejemplo, en Pandora y Last.fm

Otra de las técnicas utilizadas para la recomendación de música es la de "filtrado colaborativo" (*collaborative filtering*), consistente en comparar las preferencias de un usuario con las de otros, que se encuentran recogidas en una base de datos extensa. Cuando ambos patrones son similares el sistema puede en ese momento sugerirlas al usuario, a modo de "usuarios que descargaron esta canción también descargaron…". Este tipo de filtrado se utiliza en Last.fm y en iTunes.

Figura 5. Ejemplo de filtrado colaborativo

Ejemplo de filtrado colaborativo en last.fm

1.3. Extracción de características de bajo nivel

La extracción de características es el primer paso en el proceso de clasificación automática de música. Sin embargo, basta con escuchar la canción más sencilla para percibir la cantidad de parámetros que pueden caracterizar a una canción. El ritmo o el tono de la canción son características perceptibles, por ello, como indica [Meng et al., 2007], pueden modelarse y examinarse directamente, con la consiguiente ventaja de no tener que utilizar un clasificador o, como mucho, tener que estimarse con métodos sencillos. De hecho, los parámetros perceptibles relevantes deben ser evaluados como parte de un sistema de clasificación completa [Meng et al., 2007]. Los aspectos relativos a la extracción y selección de características de bajo nivel se desarrollarán en el Capítulo 2.

De igual forma, no todas las características requieren el mismo intervalo de reproducción para ser reconocidas y clasificadas; de manera intuitiva, una persona con condiciones normales de audición y poco tiempo de escucha discrimina la voz de un hombre de la de una mujer, sin embargo debe poseer un oído extremadamente fino y/o entrenado, además de un tiempo de escucha superior al caso anterior, para poder distinguir, por ejemplo, el conjunto de instrumentos de cuerda tocados uno detrás de otro.

Esta misma característica se repite cuando una máquina debidamente entrenada clasifica música; en otras palabras, la escala de tiempos utilizada es determinante para que pueda distinguir una característica u otra. La necesidad de trabajar con distintas escalas temporales obliga a utilizar características extraídas sobre intervalos de observación de distinta duración; sirva como ejemplo la siguiente clasificación de características en función de la escala de tiempos, extraída de [Meng et al., 2005] y que se tratará con detalle en el Capítulo 2:

- Características a corto plazo: Sólo se consideran pequeños intervalos de tiempo, por tanto no contienen información sobre cómo se estructura la secuencia musical en intervalos largos.

- Características a medio plazo: Contienen información temporal como la modulación (instrumentación).

- Características a largo plazo: Contienen información estructural, como el ritmo.

Para aprovechar al máximo la información que aportan estas características es necesario sistematizar su extracción. En este apartado se tratarán, de forma resumida, tres tipos diferentes de características: Coeficientes Cepstrales de Frecuencia en escala Mel (*Mel Frequency Cepstrum Coefficients*, MFCC en adelante), Representación de Armónicos (*Harmonic Representation,* HR en adelante) y descriptores de audio del estándar MPEG-7, que se exponen a continuación.

1.3.1. MFCC y HR: Visión general y aspectos de comparación

Para justificar la relevancia de las caracterizaciones basadas en MFCC y HR se recurrirá a conceptos puramente instrumentales, de complejidad menor que los asociados a características de alto nivel, como los artistas, que incluyen voz, polifonía, arreglos, etc.

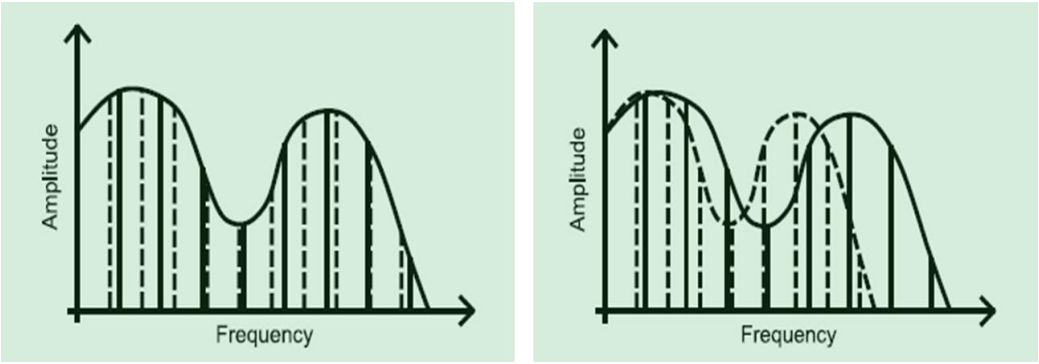
Antes de ilustrar los modelos MFCC y HR es necesario recordar qué es un armónico: un sonido no suele estar compuesto de una sinusoide a una sola frecuencia (en ese caso sería un tono puro), sino que lo conforman varias sinusoides a diferentes frecuencias. A cada una de estas frecuencias se la denomina *armónico*. La frecuencia de un armónico es un múltiplo de la fundamental, definida como la más baja dentro del espectro y que se asigna al primer armónico. La estructura de un armónico se puede caracterizar de forma muy sencilla y suficiente para los fines de este Proyecto dividiéndolo en tono y envolvente [Nielsen et al., 2007], si bien un armónico real se conforma de manera harto más complicada; volviendo a la combinación simplificada de tono y envolvente:

- El tono es lo que percibe el oído humano y su valor viene dado por la frecuencia fundamental.

- La envolvente es la modulación aplicada al tono.

Considerando las características espectrales que determinan el tono de un instrumento, se pueden establecer dos modelos contrapuestos para explicar cómo el espectro de ese instrumento en particular cambia para tonos diferentes, según se expone en [Nielsen et al., 2007]: El primer modelo establece que la envolvente del espectro permanece constante para un tono puro, mientras que el segundo afirma que es la relación entre la amplitud de los armónicos lo que permanece constante. Estos dos modelos están asociados a dos conjuntos de características: Mel Frequency Cepstrum Coefficients (MFCC) y Harmonic Representation (HR). Estos dos modelos son excluyentes entre sí. Sin embargo, tras los experimentos realizados en [Nielsen et al., 2007], [Ahrendt, P. et al., 2004] y [H.-Gook, K. & Sikora, T., 2004], se concluye que, en general, con los MFCC se obtienen mejores resultados en la clasificación y modelado de instrumentos. Así y todo, ninguno de los dos modelos es capaz de representar correctamente los datos, y tan sólo en el caso trivial de encontrarse con un tono puro ambos modelos pueden equipararse en la consecución de resultados.

Figura 6. Distribución de envolvente y armónicos en los modelos MFCC y HR



6.a) MFCC 6.b) HR

Distribución de envolvente y armónicos en los modelos MFCC y HR. Ambas figuras representan dos notas con diferentes frecuencias fundamentales. En 6.a), según dicta el modelo que da lugar a los MFCC, la envolvente permanece constante, por tanto son las amplitudes de los armónicos las que cambian (los armónicos de una nota se representan punteados y los de la otra en línea continua). En el modelo HR, (figura 6.b)), sin embargo, al mantenerse constante la relación entre armónicos, varía la envolvente, pues los armónicos para cada nota se sitúan en diferentes valores de frecuencia. Fuente: [Nielsen et al., 2007].

De acuerdo a esta aproximación, si dos instrumentos tocan la misma nota el tono será el mismo; es la envolvente, asociada al timbre, la que diferenciará ambos sonidos. El tono cambia para distintas notas, pero en este caso cómo cambia la envolvente debe explicarse con un modelo más sofisticado; este será el propósito del modelo basado en MFCC.

1.3.1.1. Coeficientes Cepstrales en Escala Mel (MFCC)

Según este modelo, desarrollado originalmente para el procesado de voz, y teniendo como base el modelo simplificado expuesto anteriormente, la envolvente del espectro de un instrumento no varía con el tono. En consecuencia, cuando el tono cambia, la amplitud de cada armónico del sonido sufre modificaciones.

Para el cómputo de los MFCC, [Logan, 2000], es preciso seleccionar una pequeña ventana de los datos de audio y aplicar la Transformada Discreta de Fourier (DFT) para obtener las bandas de frecuencia presentes en los datos enventanados y el nivel de energía existente en cada una de ellas [Oppenheim y Schafer, 1989].

El siguiente paso adecúa los resultados anteriores a las capacidades del oído humano, cuyo comportamiento se esbozará de forma breve con las ideas recogidas de [Jurafsky y Martin, 2008]:

- La sensibilidad del oído humano varía en función de la banda de frecuencia; en frecuencias altas, alrededor de los 1000 Hz, es menos sensible.

- El comportamiento del oído humano frente al nivel de señal puede modelarse como logarítmico, es decir, las pequeñas variaciones en bajas frecuencias son mejor percibidas que a altas frecuencias.

Para modelar este comportamiento se dispone un banco de filtros en escala Mel, con el que se filtrará el módulo de la DFT anterior. A continuación se calculan los logaritmos de estas salidas, ateniéndose al comportamiento logarítmico del oído mencionado anteriormente. De esta forma se cifra la energía de la señal de entrada en diferentes bandas de frecuencia, cuyas frecuencias centrales se aproximan a las de la escala Mel[1]. Finalmente, se aplica la Transformada Discreta del Coseno (DCT).

Los MFCC, por tanto, contienen información sobre la forma de la envolvente del espectro; si la envolvente fuera constante, los MFCC extraídos de diferentes ventanas del mismo instrumento deberían ser similares, incluso si correspondieran a diferentes notas, en consonancia con la formulación del modelo.

1.3.1.2. Representación Basada en Armónicos (HR)

Este modelo, desarrollado en [Nielsen et al., 2007], asume la hipótesis de que es la amplitud de los diferentes armónicos la que permanece constante, por tanto, si se producen cambios en el tono la forma de la envolvente cambia.

Si se supone válido el modelo, la representación de los datos se obtendría de forma sumamente sencilla, pues si el tono es conocido bastaría con estimar las amplitudes de los armónicos y hallar los HR mediante el cálculo de la relación entre ellas. Si el tono no se conoce habrá que estimarlo.

1.3.2. Estándar MPEG-7

MPEG-7 surgió a finales de los años 90 fruto del trabajo del Grupo de Expertos de Imágenes en Movimiento (*Motion Pictures Experts Group*, MPEG en adelante) de la Organización Internacional para la Estandarización (Internacional Organization for Standardization, ISO en adelante) y proporciona un completo conjunto de herramientas estandarizadas para describir contenido multimedia. En lo concerniente al audio, de cuyo tratamiento versa este Proyecto, posee potencialidad para crear descripciones independientes de la forma en que el contenido esté codificado y almacenado.

[1] Escala perceptual del tono. Se toma como referencia entre la frecuencia y esta escala la paridad 1000 Hz-1000 mels. El mapeado entre ambas es lineal por debajo de 1000 Hz y logarítmico por encima de ellos, empleando la fórmula m = 1127ln [1+ (f/700)], donde m es el resultado en mels y f la frecuencia en Hz.

El aspecto de MPEG-7 que tiene relación con la extracción de características y por el cual se incluye en este apartado es el de los descriptores de bajo nivel, que serán resumidos a continuación. No obstante, y tratándose de un estándar descrito en varios niveles, se ofrece una breve información adicional con el objetivo de contextualizar la información de interés.

1.3.2.1. Elementos constitutivos de MPEG-7

Los principales elementos relativos a audio del estándar MPEG-7 son [Kim et al., 2005]:

- Descriptores (*Descriptors*, D): Definen las sintaxis y la semántica de los vectores de características de audio y sus elementos. Los descriptores relacionan una característica con un conjunto de valores.

- Esquemas de Descripción (*Description Schemes*, DS): Especifican la estructura y la semántica de las relaciones entre Descriptores, y ocasionalmente las existentes entre Esquemas de Descripción.

- Lenguaje de Definición de la Descripción (*Description Definition Language*, DDL): Define la sintaxis de herramientas de descripción de herramientas de MPEG-7 nuevas o ya existentes.

- Representación en Código Binario de Descriptores o Esquemas de Descripción: Permite el almacenamiento, transmisión y multiplexación eficientes de descriptores y esquemas de descripción, sincronización de descriptores y contenido, etc.

1.3.2.2. Descriptores

Las descripciones sobre el contenido de MPEG-7 pueden versar sobre distintos aspectos del contenido, tales como información concerniente a la creación y producción, derechos de autor, características del material, interacción con el usuario, etc.

Los descriptores de bajo nivel merecen una mención especial. Hay 17 descriptores temporales y espectrales de gran versatilidad, cuya extracción a partir del audio puede automatizarse y muestran la variación de las propiedades de éste en el tiempo o la frecuencia.

Tomando como base estos descriptores será factible analizar las similitudes entre los distintos archivos de audio y/o las similitudes dentro de un mismo archivo. Por otra parte, este tratamiento también provee de una base de representación para la clasificación de audio, directamente extraíble de los archivos.

Los 17 descriptores de bajo nivel que define MPEG-7 se agrupan de la siguiente forma:

- Descriptores Básicos: Directamente relacionados con la forma de la señal de audio en el dominio del tiempo, como la forma de onda o la energía de la señal.

- Descriptores Espectrales Básicos: Cuatro descriptores derivados del análisis tiempo-frecuencia de una señal de audio que ofrecen información sobre la envolvente, los centroides, la dispersión y la variación en la forma de la señal (“*flatness”*).

- Descriptores de Parámetros de la Señal: Son dos parámetros aplicados a señales periódicas o cuasi-periódicas. Describen la frecuencia fundamental de la señal de audio.

- Descriptores Temporales del Timbre: Se extraen de la envolvente de la señal en el dominio temporal. Dado que la envolvente describe los cambios de energía de la señal, se utilizan para describir características temporales de segmentos de sonido.

- Descriptores Espectrales del Timbre: Descriptores consistentes en características espectrales en un espacio de frecuencia lineal que se aplica a la percepción del timbre musical, especialmente a la aspereza del sonido.

- Descriptores de Base Espectral: Estos dos descriptores proyectan datos de un espacio espectral de alta dimensión a uno de baja dimensión para ayudar a la compactación y al reconocimiento.

MPEG-7 no especifica cómo han de extraerse estos parámetros, pero existen librerías de libre distribución que podrían utilizarse para disponer de ellos.

Como compendio de este apartado, es necesario tener la visión de un clip de sonido compuesto por series de tiempo multivariantes de características que deben extraerse, pues son determinantes para una clasificación posterior de la canción; la siguiente fase será combinar esta información y en función de ella etiquetar el clip de sonido completo [Meng et al., 2007].

1.4. Integración temporal

En el apartado anterior se ha comentado que previamente a la clasificación es recomendable extraer distintas características de los archivos de audio disponibles a la escala temporal más apropiada. La idea general [Meng et al., 2007] es procesar ventanas temporales de tamaño fijo de la señal de audio digitalizada con un algoritmo que pueda extraer información relevante de la muestra musical. El tamaño de la ventana determina la escala de tiempos de la característica. Las características están ideadas para representar aspectos de la música tales como el tono, la instrumentación, la armonía o el ritmo.

La mayoría de las características directamente extraídas del audio que se utilizarán en este Proyecto son de corto plazo (cf. Clasificación de Características, apartado 1.3), para las cuales se emplean ventanas temporales de una duración de entre 20 y 40 ms. En la extracción de estas características se asume el hecho de que el fragmento de señal enventanado es estacionario, por lo tanto el tamaño de la ventana debe ser pequeño; estas características poseen por lo general estructura repetitiva.

Tras la extracción de características de un fragmento musical utilizando pequeñas ventanas temporales se posee la información necesaria para el reconocimiento, pero carente de una organización que la haga comprensible para una máquina. En el proceso de integración temporal se tratará este punto, utilizando la técnica de combinar los vectores de características de cada ventana temporal obtenidos anteriormente en un vector de características único, de forma que toda la información temporal relevante relativa a esa ventana quede recogida de forma compacta y aprovechable para futuros procesos, como analizan de forma extensa [Meng 2006] y [Meng et al., 2007] y como puede observarse en la Figura 7. Cabe destacar que esta nueva generación de características no captura necesariamente aquéllas fácilmente perceptibles, sino información útil para el clasificador posterior. En otras palabras, las características que se obtienen con un mayor horizonte temporal pueden carecer de interpretación física.

Figura 7. Esquema gráfico de la integración temporal

Esquema gráfico de la integración temporal. En primer lugar se extraen las características de los fragmentos de señal enventanados. Posteriormente estos vectores de características se combinan en otros, intentando minimizar la pérdida de información. Tras las etapas de integración, se obtiene un solo vector que contiene la información sobre las características de alto nivel de la canción. Los cuatro niveles, del superior al inferior, son el nivel de audio, bajo nivel, nivel de integración temporal y nivel semántico.

En la integración temporal se aúna la información de ventanas de tiempo de menor tamaño en un vector asociado a una ventana de mayor duración. Si la escala de tiempos es la adecuada, se asume una pérdida mínima de la información temporal importante para la clasificación del género, el artista o el instrumento musical.

La integración temporal de características es una técnica muy común y para la que se pueden aplicar varios modelos estadísticos, desde un modelo en el cual las características extraídas siguen una distribución gaussiana, justificando así el uso de elementos estadísticos básicos como la media y la matriz de covarianzas para realizar la estimación, hasta modelos de mayor complejidad. En este Proyecto se hará referencia a dos, los Coeficientes Autorregresivos (AR) y el Periodograma.

El primero puede utilizarse como alternativa al conjunto de media y varianza. La principal ventaja del modelo autorregresivo, como muestra [Meng et al., 2005], es su habilidad para modelar tanto elementos temporales dinámicos como las dependencias entre las dimensiones de las características a corto plazo. De hecho, el modelo es una generalización del modelo de integración temporal de características utilizando la media y la varianza. En el segundo modelo, a partir de la estimación de la secuencia de autocorrelación de un conjunto de datos se obtiene una estimación del espectro de potencia [Hayes, 1996]. Su uso está muy extendido cuando se persigue encontrar patrones de periodicidad basándose en esta estimación de la densidad espectral. Ambos son ampliamente utilizados y serán descritos de forma resumida en el siguiente apartado. Asimismo, su efectividad será analizada en los experimentos del Capítulo 4.

1.4.1. Técnicas de integración temporal

Las técnicas de integración temporal que se van a presentar se engloban dentro de los métodos de estimación de la densidad temporal, que se dividen en

- Métodos paramétricos: Son aquellos en los que la densidad espectral de potencia se estima a partir de una señal que se supone salida de un sistema lineal con adición de ruido blanco o de un determinado modelo de proceso estocástico. Para hallar la densidad espectral de potencia se estiman los coeficientes del sistema lineal que hipotéticamente genera la señal. Están indicados cuando la longitud de los datos es relativamente corta. Un ejemplo de estos métodos es el modelo autorregresivo.

- Métodos no paramétricos: Se utilizan cuando no se tiene un conocimiento a priori de la estructura de la señal, por lo que la densidad espectral de potencia se estima directamente a partir de la misma señal. Para ello se limitan áreas en el espacio de muestras. Si bien la integración temporal no está directamente ligada a la estimación espectral de potencia, es una valiosa herramienta en su consecución. A este grupo pertenece el periodograma.

1.4.1.1. Periodograma

El periodograma es un método no paramétrico consistente en calcular el espectro de potencia de un proceso aleatorio estacionario en sentido amplio en el que se hace uso de la transformada de Fourier. El periodograma fue concebido para el estudio de periodicidades, por lo que su uso se postula lógico en aplicaciones de voz, dadas las características de ésta.

Para la construcción del periodograma a partir de una serie temporal se deben ejecutar los siguientes pasos:

- Estimar la secuencia de autocorrelación de los datos reales, que en este caso serán los MFCC extraídos de los archivos de audio.

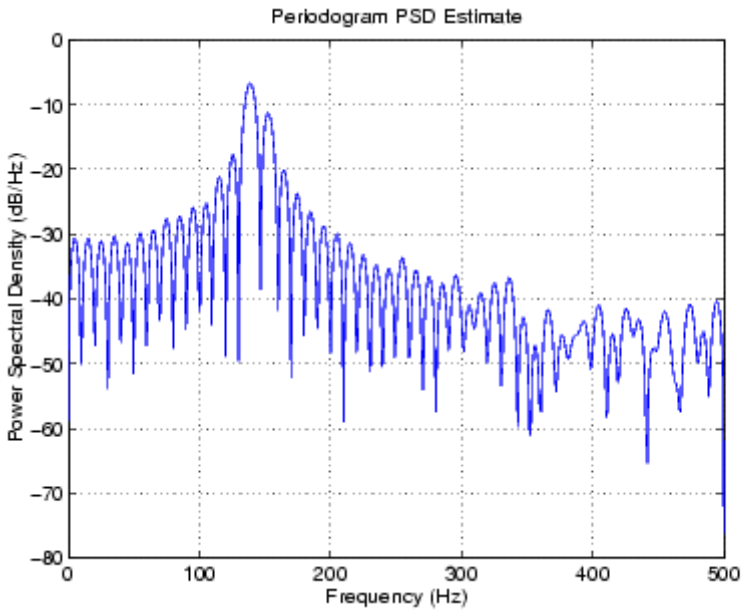
- Enventanar la secuencia de autocorrelación. Este paso se realiza comúnmente con la ventana de Hamming, pero se puede llevar a cabo con cualquier otro tipo de ventana.

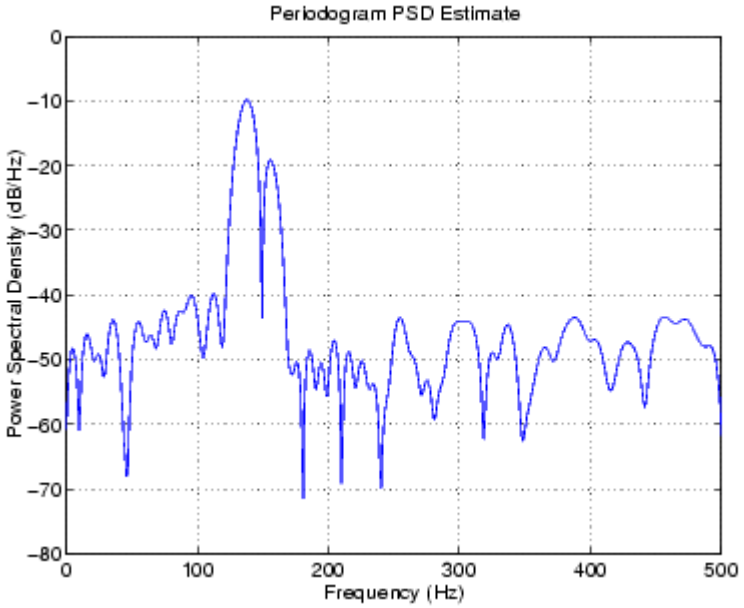
- Hallar la transformada de Fourier en tiempo discreto de los fragmentos de secuencia de autocorrelación enventanados.

- Tomar el cuadrado del valor absoluto de la transformada de Fourier calculada.

Con el fin de ilustrar estos conceptos y la forma del periodograma en función del enventanado se aplicarán dos ventanas temporales distintas sobre la misma señal. Los resultados se muestran en la Figura 8.

Figura 8. Ejemplos de enventanado de periodogramas





Ejemplos de periodograma utilizando ventana rectangular y de Hamming, respectivamente, sobre una misma señal. Comparando ambas imágenes se capta perfectamente el compromiso anchura de lóbulo principal – altura de lóbulos secundarios. Por otra parte, el hecho de utilizar ventanas no rectangulares incide sobre la potencia media de la señal, pues algunas de las muestras se atenúan cuando se multiplican por la ventana; para solventar este inconveniente existen versiones mejoradas del periodograma en las cuales se normaliza la ventana para que tenga una energía media igual a la unidad; de esta manera, la elección de ventana no afecta a la energía media de la señal.
(Fuente: http://www.mathworks.com/access/helpdesk_r13/help/toolbox/signal/spectra8.html).

Para medir la eficacia del periodograma como método de estimación de la densidad espectral se utilizan los siguientes parámetros: leakage, resolución, sesgo y varianza.

- Leakage: Es un efecto del análisis frecuencial de señales de dimensión finita en el que una parte de la energía del espectro de la señal original se desplaza hacia otras frecuencias. Una señal enventanada o truncada posee un continuo de energía trasladado por los alrededores de la frecuencia central de la ventana. Este efecto tan solo depende de la longitud de los datos, y no del hecho de que el periodograma se construya sobre la base de un número finito de muestras en frecuencia.

- Resolución: se refiere a la capacidad para discriminar características espectrales. Si dos sinusoides están relativamente cercanas entre sí en frecuencia, es necesario que la diferencia entre las dos frecuencias sea mayor que el ancho del lóbulo principal de cada una de las dos sinusoides. Este ancho se define como el ancho del lóbulo principal en el punto en el que la energía es igual a la mitad de la máxima energía de dicho lóbulo, esto es, cuando caen 3 dB desde punto de máxima energía.

- Sesgo: diferencia entre el valor esperado de un estimador y el verdadero valor del parámetro que está siendo estimado. El sesgo aporta una idea de error fijo.

- Varianza: mide cuánto variarán cada una de las estimaciones del algoritmo de aprendizaje entre sí. En contraposición al sesgo, la varianza aporta la noción de error variable.

En el Capítulo 2 se ahondará en las propiedades del periodograma; asimismo se proporcionará un desarrollo teórico más extenso de este método de estimación del espectro de potencia.

1.4.1.2. Modelo Autorregresivo

El modelo autorregresivo, contrariamente al periodograma, es un modelo paramétrico pero con mayor simplicidad conceptual que éste.

Las observaciones de una serie temporal están asociadas a los diferentes valores que alcanza una magnitud que varía con el tiempo. Posteriormente, estos valores se almacenan de forma cronológica, es decir, $(x_1, x_2, \ldots, x_N)$.

El ajuste *autorregresivo de orden p*, AR(*p*), se basa en estimar la observación *i*-ésima como una combinación lineal de las *p* observaciones anteriores más un error aleatorio, de la forma

$$x_i = \phi_0 + \sum_{j=1}^{p} \phi_j x_{i-j} + e_i \qquad (1.1)$$

En este modelo se realizan las siguientes asunciones [Peña Sánchez de Rivera, D., 1992]:

- El proceso es estacionario y ergódico:

  - Por estacionareidad, la media y la varianza de las observaciones son constantes y finitas, y la covarianza entre pares de ellas depende tan sólo de su separación temporal.

  - Por ergodicidad, la media y la varianza de la distribución coinciden con la media y la varianza de las muestras.

- Los residuos e_i son independientes e idénticamente distribuidos, haciéndolo como gaussianas de media nula y varianza σ^2 desconocida. Por otra parte, son también independientes de las observaciones x_i del proceso.

A partir de los datos observados x_i, el modelo AR(p) tiene p+2 parámetros que debe estimar:

- Los p+1 coeficientes autorregresivos $(\phi_0, \phi_1, \ldots, \phi_p)$

- La varianza del residuo, σ^2

El método de estimación adoptado es el de mínimos cuadrados, que consiste en calcular los parámetros autorregresivos de forma tal que se minimice el error cuadrático:

$$\text{EC} = \sum_{i=p+1}^{N} \left(x_i - \left[\phi_0 + \sum_{j=1}^{p} \phi_j x_{i-j} \right] \right)^2 \qquad (1.2)$$

Con el apoyo de la teoría de regresión lineal, se concluye que la estimación de los coeficientes autorregresivos se efectuará como sigue:

$$\hat{\boldsymbol{\varphi}} = \begin{pmatrix} \hat{\phi}_0 \\ \hat{\phi}_1 \\ \vdots \\ \hat{\phi}_p \end{pmatrix} = (X^T X)^{-1} X^T T, \qquad (1.3)$$

definiéndose las matrices X y T como

$$\mathbf{X} = \begin{pmatrix} 1 & x_p & x_{p-1} & \cdots & x_1 \\ 1 & x_{p+1} & x_p & \cdots & x_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n-1} & x_{n-2} & \cdots & x_{n-p} \end{pmatrix} \qquad (1.4)$$

$$\mathbf{T} = \begin{pmatrix} x_{p+1} \\ x_{p+2} \\ \vdots \\ x_n \end{pmatrix} \qquad (1.5)$$

En cuanto a la estimación de la varianza σ^2, se utilizará el estimador de minímo error cuadrático promedio.

$$\hat{\sigma}^2 = \frac{1}{n-p}\sum_{i=p+1}^{n}\left(x_i - \left[\hat{\phi}_0 + \sum_{j=1}^{p}\hat{\phi}_j x_{i-j}\right]\right)^2 \qquad (1.6)$$

Para concluir si el modelo AR(*p*) representa de forma aceptable la serie numérica observada se debe verificar que las hipótesis aceptadas en el modelo sean válidas.

1.5. Aprendizaje Supervisado

Para que una máquina sea capaz de clasificar correctamente es imprescindible un aprendizaje previo. La clave está en dotarla de un rango de funciones lo suficientemente amplio como para clasificar datos desconocidos de la forma más exacta posible, evitando tanto el subajuste (por falta de potencia expresiva del modelo propuesto) como el sobreajuste (clasificación tan precisa de los datos de entrenamiento que impide de todo punto la generalización). Este aprendizaje estará destinado a alcanzar de forma aceptable el objetivo de clasificar cada vector resultante de la integración temporal de acuerdo a la taxonomía de alto nivel establecida [Bishop, 1995].

Como delimitación del contenido a tratar a lo largo de este Proyecto se revisarán algunas nociones relativas al campo del aprendizaje y las redes neuronales.

Aprendizaje supervisado vs. Aprendizaje no supervisado

En este Proyecto se emplean conjuntos de datos etiquetados agrupados en pares $(\mathbf{x}_i, y_i)$:

- Los elementos $\mathbf{x}_i$ son los vectores de datos, que particularizando para este caso son cada una de las secuencias musicales procesadas, distribuidos en un espacio de entrada $\Re^N$.

- Los elementos $y_i \in \{1,...,K\}$ constituyen las etiquetas, es decir, el género/artista asignado a cada secuencia musical concreta.

Con la ejecución del algoritmo se produce una función de salida $f : \Re^N \to \Re$, de la que se espera que sea capaz de predecir la etiqueta y_j de nuevos ejemplos de $\mathbf{x}_j$.

Para que una máquina llegue a predecir correctamente es sometida a un proceso de aprendizaje, utilizando para ello un conjunto de datos de entrenamiento debidamente etiquetados. El objetivo será que la máquina ajuste sus parámetros en función de los errores cometidos, entendiéndose como error el hecho de que la máquina proponga una respuesta distinta a la etiqueta.

El entrenamiento se da por finalizado cuando la máquina ha alcanzado una tasa de aciertos óptima desde un punto de vista estadístico. A esta etapa le sucede otra de validación, en la que se verificará si efectivamente la máquina es capaz de superar con éxito pruebas con datos desconocidos. Éste es, a grandes rasgos, el fundamento de un algoritmo de aprendizaje supervisado.

Los algoritmos de aprendizaje no supervisado [Bousquet y Pérez-Cruz, 2003] se diferencian de los anteriores en que trabajan con datos sin etiquetar (tan sólo con $\{x_i\}$) y su objetivo es describir la estructura de los datos de entrada, por ejemplo su distribución.

Red neuronal

Seguidamente se definirá de forma sencilla el concepto de "red neuronal" vista como máquina autoajustable, según [Haykin, S., 1999]:

"*Una red neuronal es un procesador de distribución paralela, constituido a su vez por unidades de procesamiento simple denominadas "neuronas", que almacena conocimiento empírico y hace uso de él. Guarda parecido con el cerebro humano en dos aspectos:*

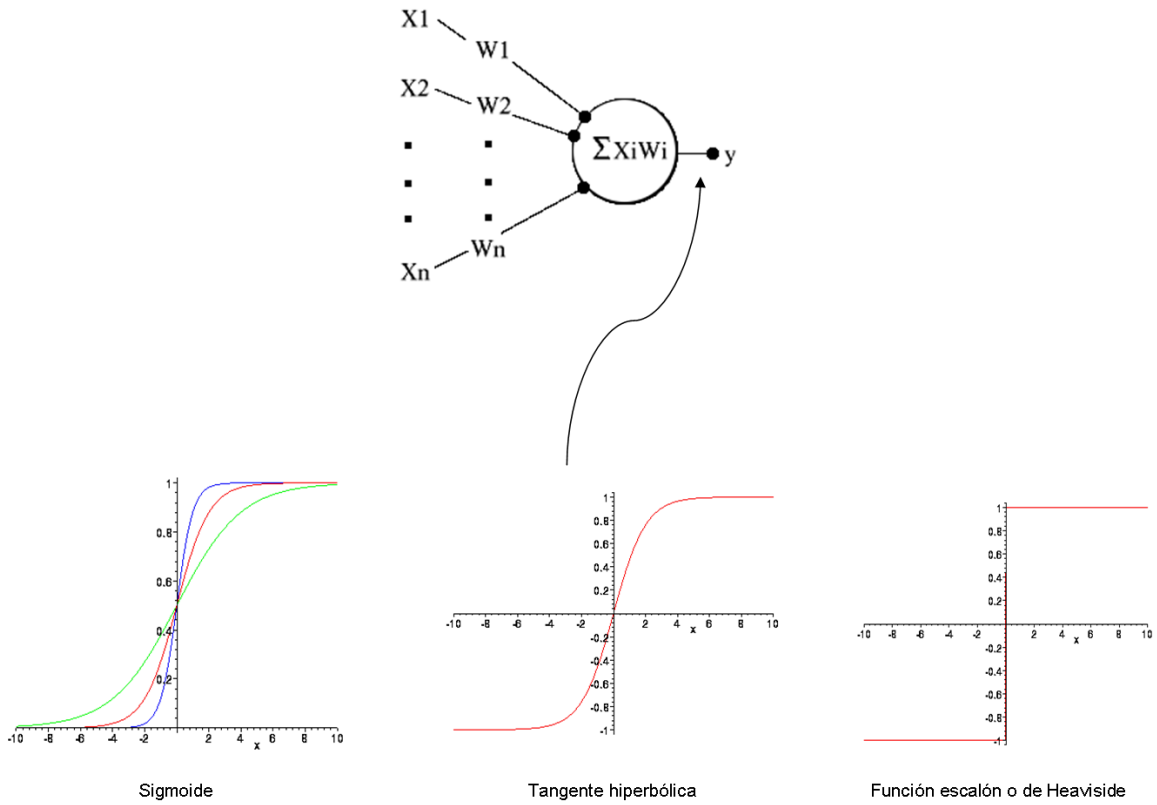
1. *La red adquiere el conocimiento del medio a través de un proceso de aprendizaje.*
2. *Las conexiones interneuronales, conocidas como pesos sinápticos, se utilizan para almacenar el conocimiento adquirido."*

A pesar de estar basada en el cerebro humano, las unidades de procesamiento distan mucho de parecerse a las neuronas de éste, tanto en número como en complejidad: Una "neurona" de *N* entradas $\{x_i\}_{i=1}^{N}$, multiplicadas cada una por un peso $\{w_i\}_{i=0}^{N}$ de la forma

$$z = w_0 + \sum_{l=1}^{L} w_i x \qquad (1.7)$$

consta de una función de activación que se aplica a *z*. Esta función de activación acota la salida y varía en función de la interpretación que se le vaya a dar a ésta. Las más utilizadas son las funciones sigmoide y tangente hiperbólica, ambas de activación blanda, en detrimento de la función de Heaviside [Bishop, 1995], de activación dura. Este proceso se esquematiza en la Figura 9.

Figura 9. Esquema de una red neuronal



Esquema de una red neuronal con indicación del punto de aplicación de la función de activación.

Métodos de núcleo

Los métodos de núcleo permiten derivar versiones no lineales de algoritmos lineales, aumentando así su capacidad expresiva. Se ahondará en este concepto en el Capítulo 3, por ser este tipo de algoritmos los empleados en la fase de clasificación de este Proyecto.

1.6. Objetivos del Proyecto de Fin de Carrera

Una vez se han presentado someramente los distintos pasos que tienen lugar en el proceso de reconocimiento y clasificación automática de música, se está en disposición de presentar los objetivos perseguidos con la realización del presente Proyecto de Fin de Carrera.

El objetivo del presente Proyecto es estudiar la influencia de la ventana temporal, analizando qué horizontes temporales representan de forma más precisa las distintas características de alto nivel de las señales de audio en los experimentos de clasificación musical, así como determinar el valor de dicha ventana en algunos casos concretos. En este primer Capítulo se ha pretendido describir las fases de las que consta un clasificador de información musical, haciendo especial hincapié en aquellas que, a tenor de los experimentos previos realizados en la clasificación de canciones, se utilizarán en la parte práctica.

El Capítulo 2 se dedicará a describir de forma detallada la extracción de características, integración temporal y aprendizaje máquina. El Capítulo 3 se centrará en la clasificación máquina, primero ofreciendo una visión general para después exponer los métodos utilizados en el Proyecto. El cuarto capítulo describirá los experimentos realizados y mostrará los resultados y conclusiones. Por último, el quinto capítulo listará las conclusiones del Proyecto y las líneas futuras que sugiere el trabajo realizado.

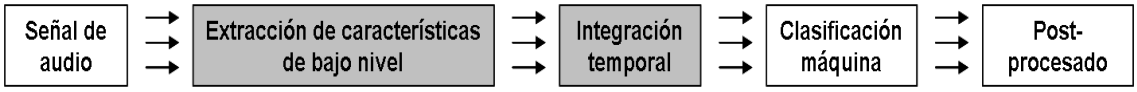
CAPÍTULO 2. Extracción de Características de bajo nivel e Integración Temporal

2.1. Introducción

En este capítulo se tratarán los dos primeros pasos del proceso de inferencia de características de alto nivel a partir de aquéllas de bajo nivel. Como puede observarse en los cuadros sombreados de la Figura 10, la primera etapa consiste en la extracción de características de bajo nivel directamente del audio, mientras que en la segunda estos vectores de características de bajo nivel se integran en uno solo, de forma que la señal de audio de la que se ha partido quede convenientemente descrita a partir de ellas. Si la extracción de características de bajo nivel se ha efectuado de forma correcta, la combinación en este vector de horizonte temporal mayor se efectuará con la pérdida mínima de información, comparándose con la contenida en los datos originales.

Al tratar dos etapas, el capítulo puede dividirse de forma natural en dos apartados, el primero la extracción de características y el segundo la integración temporal. Los conceptos que se tratarán ya fueron introducidos y brevemente explicados en el Capítulo 1, por tanto este capítulo sirve como extensión de ellos.

Figura 10. Proceso de clasificación general: Extracción de características e integración



Situación de las etapas de Extracción de características de bajo nivel e Integración temporal dentro del diagrama de bloques de clasificación de alto nivel de señales de audio.

2.2. Contexto de la Extracción de Características de Audio

Como ya se adelantó en el primer capítulo, una de las cuestiones determinantes que se deben plantear en la extracción de características es el salto semántico, puesto que el éxito en la inferencia de las características de alto nivel depende directamente de las características de bajo nivel extraídas. En ese mismo capítulo se expusieron de forma preliminar tres caracterizaciones: MFCC, HR y, de modo ilustrativo, las características de bajo nivel de las que hace uso MPEG-7; asimismo, y a la vista de los resultados en [Ahrendt et al., 2004], se concluye que con los MFCC se extraen las características más relevantes de cara a la posterior clasificación, si bien este resultado queda condicionado por la naturaleza de la tarea.

Con el fin de contextualizar las características de bajo nivel y los métodos de integración temporal utilizados en este Proyecto, se va a detallar la extracción de características a corto, medio y largo plazo, que ya fue introducida en el Capítulo 1 y que puede encontrarse íntegra en [Meng et al., 2005]:

- Características a corto plazo: Son las características más comunes en la literatura. Se consideran pequeños intervalos de tiempo, entre los 20 y los 40 ms [Meng et al., 2007], de forma que el fragmento enventanado sea estacionario. Típicamente se utiliza una ventana temporal de alrededor de 30 ms y un solapamiento entre ventanas temporales de 10 ms [Meng et al., 2005]. En consecuencia, no aporta información sobre cómo se estructura la secuencia musical en intervalos largos. Para obtenerlas, el procedimiento más utilizado es una transformación al dominio frecuencial [Meng et al., 2007]**.** Es en este tipo de características donde se encuadran los MFCC.

  En este Proyecto se han empleado los seis primeros MFCC de cada archivo de audio, demostrándose la suficiencia de esta cantidad para los propósitos que aquí se persiguen en [Ahrendt et al., 2004], [Arenas-García et al., 2006], [Meng et al., 2005] y [Meng, 2006]. El MFCC de menor orden contiene información sobre las variaciones suaves en la envolvente espectral; a medida que aumenta el orden se registran las variaciones más rápidas. Al hablar de orden hay que remitirse al último paso del proceso mediante el cual se obtienen los MFCC y que viene detallado en el Subapartado 2.2.1. En la extracción de los MFCC es necesario contar con el solapamiento de la ventana temporal. Lógicamente, cuanto mayor sea este solapamiento más suaves serán las variaciones entre los MFCC, sin embargo se debe tener en cuenta que el solapamiento en los MFCC debe evitarse. El primer coeficiente indica la intensidad del sonido, el volumen de la grabación. Para la correcta extracción de los MFCC se impone que el volumen de la grabación sea medianamente uniforme en los distintos archivos de audio.

- Características a medio plazo: Contienen información temporal como la modulación (instrumentación). En este caso, los valores típicos de ventana temporal estarán entre medio y varios segundos, utilizando solapamientos del 50% [Meng et al., 2005]. Entre estas características se encuentran la media y la varianza de los MFCC, el modelo autorregresivo AR y el periodograma.

  La media y la varianza de los MFCC son ampliamente utilizadas en la integración temporal debido a la facilidad de su obtención. El modelo autorregresivo, que fue expuesto de forma somera en el Capítulo 1 y será ampliado en este, es uno de los métodos utilizados en los experimentos de este Proyecto para la integración temporal de los MFCC.

Otras características a medio plazo que no se desarrollarán en este Proyecto y que pueden encontrarse en [Meng et al., 2007] son dos extraídas directamente de la señal de audio: la tasa de cruces por cero (*High Zero-Crossing Rate Ratio*, HZCRR) y la tasa de energía a corto plazo (*Low Short-Time Energy Ratio*, LSTER).

- Características a largo plazo: Contienen información estructural, como el ritmo. Los valores típicos de ventana temporal y de solapamiento entre ventanas temporales serán de alrededor de 10 y 5 segundos, respectivamente [Meng et al., 2005]. Muchas características incluidas aquí son características extraídas a escalas temporales menores integradas temporalmente, formando así un vector mayor.

A continuación se va a profundizar en los dos aspectos centrales de este capítulo: la extracción de características a corto plazo, más concretamente de los MFCC, coeficientes que han sido utilizados en la realización de este Proyecto, y la integración temporal de estas características a corto plazo, mediante la cual se obtendrán características a medio plazo, ampliando así la información relativa a la señal de audio que puede ser aplicada en la etapa posterior de clasificación máquina (cf. Figura 10).

2.3. Extracción de Características a Corto Plazo: MFCC

El proceso de extracción de los MFCC, cuya secuencia se ilustra en la Figura 11, fue expuesto de forma somera en el Capítulo 1. Seguidamente se profundizará en este proceso, exponiéndolo con un tratamiento más técnico. El procedimiento descrito a continuación se ha tomado de [Sigurdsson et al., 2006].

Figura 11. Proceso de extracción de MFCC

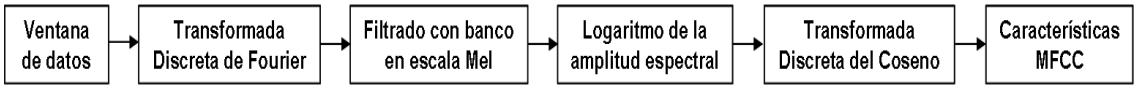


Diagrama de bloques del proceso de extracción de MFCC a partir de la ventana de datos.

2.3.1. Transformada Discreta de Fourier

El primer paso dentro de este proceso consiste en aplicar la Transformada Discreta de Fourier a los fragmentos de datos enventanados. Como ya se adelantó en las propiedades de las características a corto plazo en el Subapartado 2.2.1., el tamaño de la ventana temporal debe ser tal que los datos dentro de ella conserven la condición de estacionariedad. En cada ventana temporal se aplicará dicha transformada,

$$X(k) = \sum_{n=0}^{N-1} w(n)x(n)\exp(-j2\pi kn/N), \text{ con } k = 0,\ldots,N-1 \qquad (2.1.)$$

donde *x(n)* son los datos en tiempo discreto y *w(n)* la ventana temporal de tamaño *N*; sobre el tipo de ventana temporal es conveniente apuntar que se puede utilizar cualquier tipo, sin embargo, varios puntos de la bibliografía en los que se incluye esta implementación de los MFCC, como por ejemplo [Meng, 2006] o [Sigurdsson et al., 2006], asumen como ventana temporal la de Hamming, definida como *w(n)* = $0.54-0.46\cos(\pi n/N)$, y ésta será la utilizada a lo largo del desarrollo teórico y experimental. Por último, el parámetro *k* indexa los valores de frecuencia $f(k) = kf_s/N$, siendo f_s la frecuencia de muestreo en Hz de la señal original.

2.3.2. Filtrado en escala Mel

El siguiente paso consiste en filtrar el valor absoluto de *X(k)* empleando un banco de filtros de Mel *H(k,m)*; la conveniencia de esta escala, como ya se expuso en el Capítulo 1, consiste básicamente en enfatizar las frecuencias más bajas, en consonancia con el comportamiento del oído humano. Este banco de filtros se conforma como un conjunto de *M (M<<N)* filtros triangulares cuyas frecuencias centrales $f_c(m)$, con *m* = 1, 2,…, *M*, coinciden aproximadamente con las resultantes al operar con la escala Mel, cuya fórmula de conversión es, como se recordará del primer capítulo, m = 1127ln [1+ (f/700)], donde m es el resultado en mels y f la frecuencia en Hz. Nótese que tras la transformación a escala Mel, las frecuencias quedan escaladas de forma logarítmica. El banco de filtros Mel así descrito se formula como sigue:

$$H(k,m) = \begin{cases} 0 & para \quad f(k) < f_c(m-1) \\ \dfrac{f(k)-f_c(m-1)}{f_c(m)-f_c(m-1)} & para \quad f_c(m-1) \le f(k) \le f_c(m) \\ \dfrac{f(k)-f_c(m+1)}{f_c(m)-f_c(m-1)} & para \quad f_c(m) \le f(k) \le f_c(m+1) \\ 0 & para \quad f(k) \ge f_c(m+1) \end{cases},$$

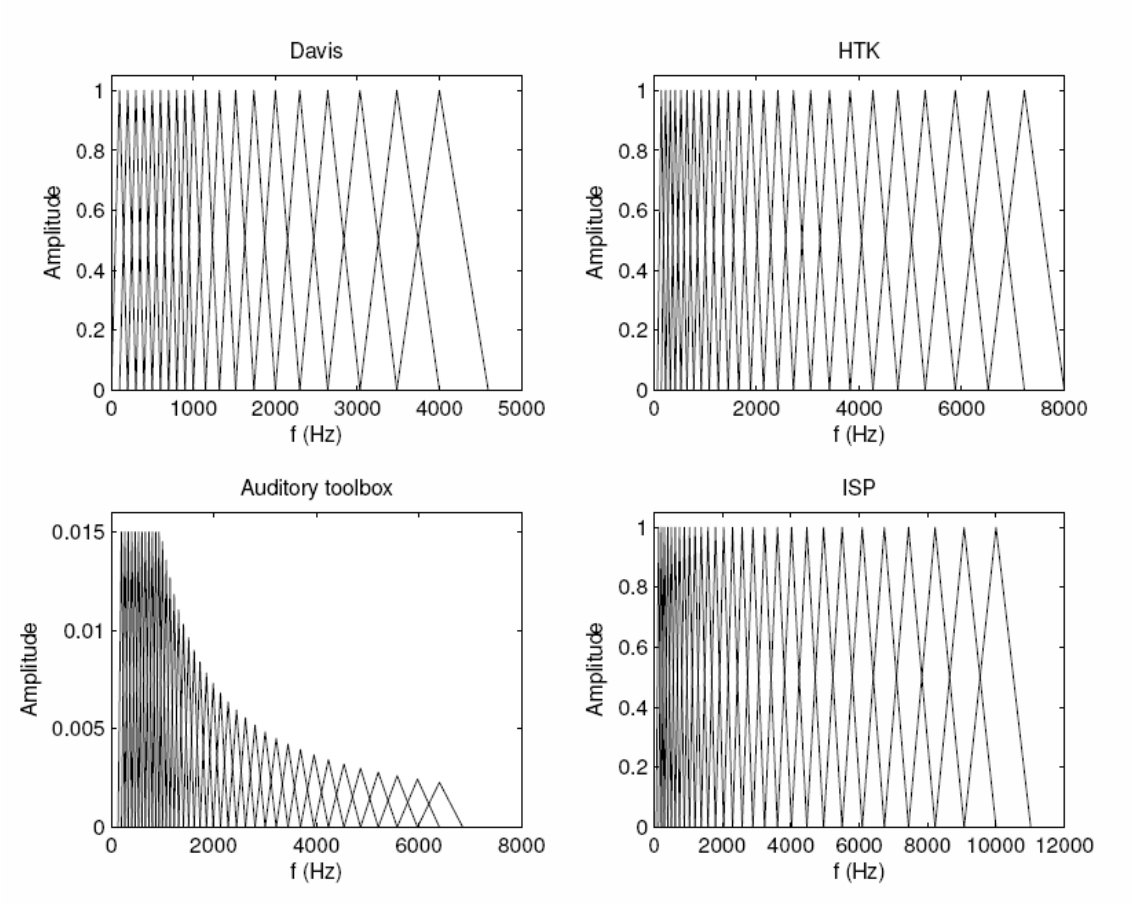
donde *m* indica el filtro utilizado y *k* corresponde al valor de la frecuencia según la descripción previa.

La operación de filtrado puede expresarse de forma compacta como

$$X(m) = \sum_{k=0}^{N-1} |X(k)| \cdot H(k,m). \qquad (2.2)$$

La disposición del banco de filtros de Mel puede variar en función de las distintas implementaciones existentes. Ejemplos de distintos bancos de filtros de Mel y las implementaciones en las que se utilizan pueden apreciarse en la Figura 12:

Figura 12. Implementaciones de bancos de filtros Mel



Distintas implementaciones de bancos de filtros Mel. Fuente: [Sigurdsson et al., 2006].

2.3.3. Logaritmo de la amplitud espectral y Transformada Discreta del Coseno

Siguiendo el proceso enunciado en la Figura 11 se toma el logaritmo de las salidas de los filtros en escala Mel.

$$X'(m) = \ln\left(\sum_{k=0}^{N-1} |X(k)| \cdot H(k,m)\right) \tag{2.3}$$

y finalmente se obtienen los MFCC mediante la aplicación de la Transformada Discreta del Coseno (DCT) a $X'(m)$:

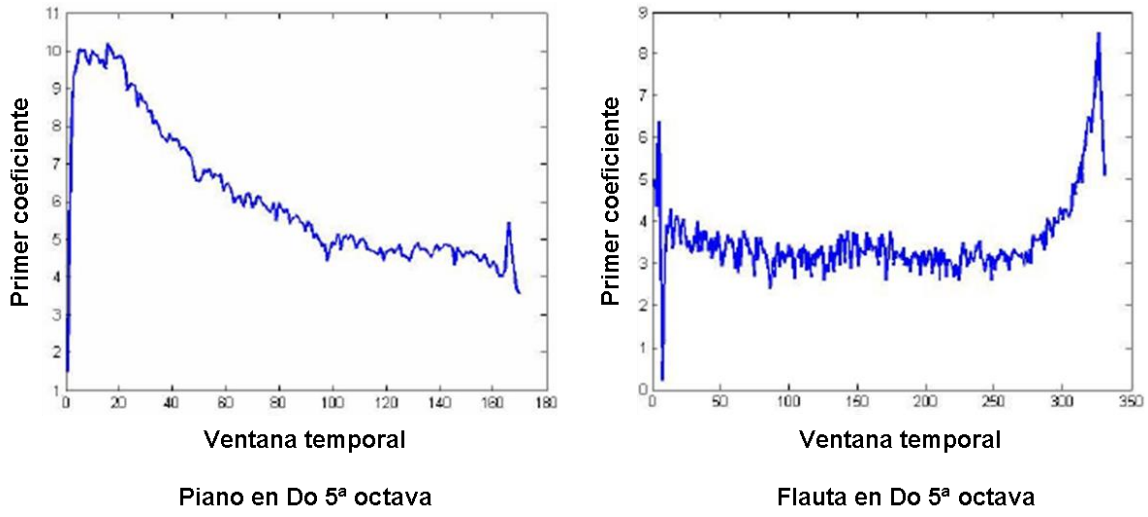
$$c(l) = \sum_{m=1}^{M} X'(m) \cos\left(l \frac{\pi}{M}\left(m - \frac{1}{2}\right)\right), \tag{2.4}$$

donde $l = 1,2,\ldots, M$ y $c(l)$ es el MFCC l-ésimo.

La DCT actúa como una suerte de Análisis de Componentes Principales (PCA). Entrando en poco detalle, el objetivo del PCA es reducir la dimensionalidad de los datos mediante proyección; de esta forma se crea un nuevo sistema de coordenadas en el que las varianzas mayores coinciden con los ejes, es decir, se consigue una mayor ortogonalización de los datos. La DCT también actúa en el sentido de reducir la dimensionalidad de los datos, pero el resultado es una serie de valores incorrelacionados para cada ventana temporal, los MFCC [Loughran et al., 2008].

Los MFCC describen cada una de las ventanas frecuenciales de la señal original y, volviendo a la introducción del Capítulo 1, están muy relacionados con la envolvente de la señal de audio. Esta envolvente permite identificar los distintos archivos de audio, pues diferentes notas provenientes de la misma fuente poseen envolventes parecidas. Como ejemplo, en la Figura 13 se muestra el segundo MFCC para un piano y para una flauta, ambos en Do 5ª octava.

Figura 13. Ejemplo de primer coeficiente MFCC para piano y flauta en la misma nota



Primer MFCC para un piano y una flauta, ambos tocando la misma nota. Asumiendo que las distintas envolventes resultantes de tocar con un piano diferentes notas serán parecidas, e ídem con la flauta, puede apreciarse cómo el hecho de distinguir un piano de una flauta a partir de sus envolventes es relativamente sencillo. Fuente: [Loughran et al., 2008]

Durante la fase de entrenamiento la máquina aprenderá a relacionar los vectores de MFCC con el instrumento, el género o el artista gracias a los datos etiquetados. Sin embargo, el reconocimiento tanto en el entrenamiento como en la fase posterior de validación no se limita a un solo MFCC, ni un solo MFCC es capaz de representar de forma suficiente el fragmento musical. Será preciso 'ensamblar' todos los vectores que componen la canción y a partir de ellos extraer un vector a mayor escala temporal; así sendas fases de entrenamiento y validación se efectuarán de forma más fiable y eficiente debido a la compactación de los datos. Este procedimiento, visto de forma breve en el

primer capítulo y enunciado en la Figura 10, es precisamente el objeto del siguiente apartado.

2.4. Integración Temporal

2.4.1. ¿Qué es la Integración Temporal?

Muchas de las características de bajo nivel utilizadas en las distintas aplicaciones de clasificación musical, minería de datos y recuperación de información son a corto plazo es decir, del orden de milisegundos, mientras el horizonte temporal en el terreno de la decisión suele ser del orden de segundos, dado que es en el que se dispone de información suficiente. Si los vectores de bajo nivel se procesan adecuadamente la toma de decisiones puede efectuarse con un bajo nivel de error. Es por ello necesario realizar una integración temporal de las características a corto plazo para conseguir discernir a qué escala está la información relevante y así satisfacer los requerimientos necesarios en la decisión. Esta información relevante dependerá de la tarea concreta que se deba realizar.

El medio de salvar este salto temporal es la integración temporal, consistente en reunir en un vector único de características toda la información procedente de la evolución temporal de los coeficientes de bajo nivel. Es importante hacer notar dos puntos [Meng, 2006]:

- Las características que conforman un vector integrado a una escala temporal mayor no tienen por qué ser perceptibles o estar relacionadas físicamente con algún parámetro musical, sino que a menudo son elaboraciones realizadas a partir los vectores de características a escala temporal menor, pero a efectos de riqueza sobre información temporal el nuevo vector supera a éstos.

- El hecho de construir vectores de características a un horizonte temporal mayor implica una reducción del número de muestras, por lo que este proceso puede considerarse también desde el punto de vista de la compresión de los datos.

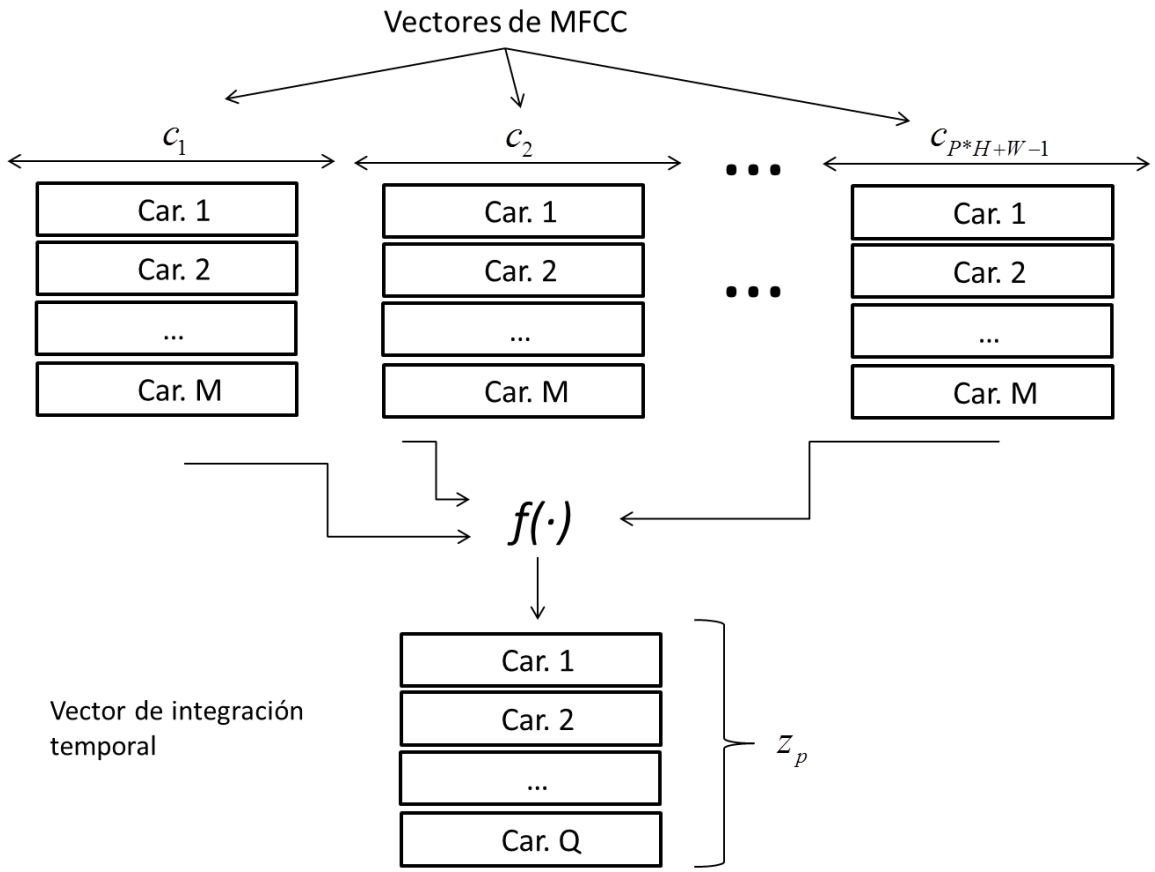
Una representación formal de la integración temporal, que puede encontrarse en [Meng et al., 2007], es la siguiente:

$$\mathbf{z}_p = f(\boldsymbol{c}_{p\cdot H}, \ldots, \boldsymbol{c}_{p\cdot H+W-1}) \qquad (2.5)$$

donde **c** son los vectores de características a corto plazo, en este caso vectores con los coeficientes MFCC *c(l)*, $l = 1, \ldots, M$, correspondientes a cada ventana de bajo nivel. W es el número de ventanas a corto plazo que componen la ventana para la integración temporal y H el solapamiento entre ventanas, ambos expresados en número de muestras; $p \in \mathrm{N}$ es el índice de tiempo discreto de la escala temporal mayor, y finalmente $f(\cdot)$ es

la función que mapea la secuencia de MFCC en el nuevo vector de características. Este proceso puede verse desarrollado gráficamente en la Figura 14.

Figura 14. Esquema de integración temporal



Ejemplo gráfico de la integración temporal. A partir de los vectores de características MFCC y mediante una función de mapeado $f(\cdot)$ se consigue un vector de integración temporal de mayor horizonte temporal en el que las características de bajo nivel extraídas anteriormente quedan debidamente compactadas.

Esta representación, vista tanto formal como gráficamente, será utilizada en las formulaciones más frecuentes de métodos de integración temporal, que serán expuestas en el siguiente apartado.

2.4.2. Métodos de Integración Temporal

El objetivo de la integración temporal es, como ya se ha mencionado, extraer las características que mejor representen a una ventana de duración media de una señal de audio, con el fin de reunirlas en un vector útil para la clasificación posterior, sin menoscabar por ello la eficiencia en la compactación de los datos; para ello existen diversos métodos, que se explicarán a continuación de forma breve con el fin de

contextualizar el método de integración temporal utilizado en este Proyecto, el modelo autorregresivo.

En la literatura consultada, [Meng, 2006], [Meng et al., 2007] y [Meng y Shawe-Taylor, 2005], los métodos de integración temporal más usuales son los que se enumeran a continuación, en orden de complejidad:

- Apilamiento
- Modelos Gausianos
  - Simple
  - Multivariante
- Coeficientes de Banco de Filtros
- Modelos Autorregresivos (AR)
  - Diagonal (DAR)
  - Multivariante (MAR)

2.4.2.1. Apilamiento

El apilamiento o *stacking* es el método más simple. Consiste en agrupar todos los vectores de características a corto plazo consecutivamente, manteniendo toda la información, esto es, sin ningún tipo de compresión de los datos. Como podrá advertirse, las dimensiones del vector final serán muy elevadas, con el consiguiente gasto de recursos.

2.4.2.2. Modelos Gausianos

El modelo gausiano asume que las muestras consecutivas de las características a corto plazo son independientes y se distribuyen de acuerdo a una distribución gausiana. Aplicando el criterio de máxima verosimilitud se obtendrá un único vector $\mathbf{z}_k$ con las estimaciones de media y varianza de las características a corto plazo:

$$\mathbf{z}_p = \begin{bmatrix} \mathbf{m}_p \\ \mathbf{v}_p \end{bmatrix}, \qquad (2.6)$$

que serán dependientes de los vectores de MFCC y su dimensión y de los tamaños de ventana y solapamiento.

A diferencia del apilamiento, el modelo gausiano sí introduce una compresión de los datos; sin embargo, asumir la independencia entre características no es correcto debido a la propia naturaleza de la música. Para solventar este problema se puede asumir un modelo gausiano multivariante, en el que se modelan las correlaciones entre características, incluyendo por tanto las covarianzas en la formulación del modelo:

$$\mathbf{z}_p = \begin{bmatrix} \mathbf{m}_p \\ triag(V_p) \end{bmatrix}, \tag{2.7}$$

donde $triag(\boldsymbol{V}_p)$ son los componentes de la matriz triangular superior que contiene las varianzas y covarianzas.

Con este último método se consigue modelar las correlaciones entre características, sin embargo, ninguno de los modelos gausianos tiene en cuenta las dependencias temporales entre los datos, sino que consideran a los vectores de MFCCs como realizaciones independientes e idénticamente distribuidas de una distribución de probabilidad.

2.4.2.3. Coeficientes de Banco de Filtros. El Periodograma.

Este método se basa en la extracción de energía de cada característica a corto plazo de forma independiente y en bandas de frecuencia específicas.

El periodograma es un método no paramétrico consistente en calcular el espectro de potencia de un proceso aleatorio estacionario en sentido amplio mediante la transformada de Fourier de la secuencia de autocorrelación, a su vez previamente estimada partiendo de datos reales. Para el análisis del periodograma es necesario escoger de forma adecuada el enventanado temporal y el grado de suavizado en función de la serie de datos, pues influyen de manera significativa en el resultado final.

El desarrollo teórico se hará partiendo de la realización de procesos estocásticos estacionarios en sentido amplio; considerando *c(l)* como el conjunto de los MFCC de una canción, con la técnica del periodograma se conseguirá la distribución espectral de las características. De este modo, y variando el tamaño de ventana para conseguir las estimaciones óptimas de las características, pueden establecerse patrones que conduzcan a la predicción de las características de alto nivel.

Siguiendo los pasos establecidos en el Capítulo 1 para construir el periodograma, en primer lugar se calculará la secuencia de autocorrelación de *c(l)*, de la forma

$$r_c(l,k) = E[c_p^{\,*}(l)c_{p-k}(l)], \tag{2.8}$$

que se estimará mediante promedio para trabajar con secuencias finitas de longitud N, de la forma

$$r_c(l,k) = \frac{1}{N}\sum_{l=1}^{N} c_p^{\,*}(l)c_{p-k}(l), \tag{2.9}$$

Este estimador es insesgado - su esperanza coincide con la función de autocorrelación que se pretende estimar -, pero no consistente, por tanto la varianza de la estimación no será nula en el caso de un número muy elevado de muestras. Para conseguir un estimador de mejores condiciones la señal $c_p(l)$ se enventanará previamente, mediante el producto $c_N(l) = v(l)c_p(l)$. Este producto introduce un factor que modifica los valores originales de la señal de muestra, denominado "factor de ventana" y denotado como *v(n)*; con la señal $c_p(l)$ limitada al intervalo [1, *N*] (ver límites del sumatorio), la secuencia de autocorrelación podrá reescribirse como sigue:

$$\hat{r}_c(l,k) = \frac{1}{N}\sum_{l=0}^{N} c_p(l)c_p^*(l-k) \tag{2.10}$$

Es en este punto donde se puede analizar una de las debilidades del periodograma: si bien es un método de cálculo simple, al valerse de la estimación a partir de los datos reales la secuencia debe ser larga para obtener una buena representación, puesto que el periodograma debería converger al verdadero valor del espectro de potencia; el hecho de que esta secuencia deba ser larga entra en contradicción con la estacionariedad, necesaria en la definición de la función de autocorrelación. Por otro lado, es sensible a la ventana utilizada, pues la elección de la misma determina el suavizado de la señal inherente al periodograma; por ejemplo, la elección de una ventana rectangular supone un lóbulo principal estrecho, pero introduce lóbulos secundarios relativamente grandes, que tienden a enmascarar componentes de banda estrecha de baja potencia. La reducción de los lóbulos secundarios ha de llevarse a cabo a expensas del lóbulo principal, que se ensanchará, lo que lleva a una pérdida de resolución. Este proceso puede verse gráficamente en el subapartado dedicado al periodograma del Capítulo 1.

Siguiendo con el desarrollo, se aplica la transformada de Fourier a esta secuencia de autocorrelación; haciendo uso del teorema de la convolución, el periodograma toma la forma:

$$\hat{P}_x(e^{j\omega}) = \sum_{n=-\infty}^{\infty}\hat{r}_x(k)e^{-j\omega n} = \frac{1}{N}\sum_{n=-\infty}^{\infty}x_N(n)x_N^*(n-k)e^{-j\omega n} = \frac{1}{N}\left|\hat{X}_N(e^{j\omega})\right|^2, \tag{2.11}$$

Como puede observarse, algunas propiedades que se le pueden atribuir al periodograma son la relativa sencillez de su desarrollo teórico y su facilidad de programación. Asimismo, es un método de fácil evaluación. Sin embargo, las dependencias anteriormente apuntadas constituyen algunos de sus inconvenientes. Conviene apuntar que a esta primera formulación se le han ido añadiendo sucesivas mejoras, con la consiguiente ganancia en prestaciones. Estas formulaciones no se tratarán en esta memoria, pero pueden consultarse en [Hayes, 1996].

Una vez desarrollado el periodograma, puede procederse a la integración temporal haciendo uso del método de Coeficientes de Banco de Filtros, que se basa en la extracción de energía de cada característica a corto plazo de forma independiente y en cuatro bandas de frecuencia específicas. Aquí la integración temporal puede expresarse de la siguiente forma:

$$\mathbf{z}_p = vec(\mathbf{P}_p\mathbf{F}), \qquad (2.12)$$

$\mathbf{F}$ es una matriz de filtros de dimensión $N \times 4$, donde $N = W/2$ o $N = (W-1)/2$ en caso de W impar, y $\mathbf{P}_p$ es una matriz de periodogramas $D \times N$, donde D es la dimensión del vector de características a corto plazo.

Con el uso de este método se reflejará la evolución temporal de las características, reflejo que no ocurre con la correlación entre las componentes.

2.4.2.4. Modelos Autorregresivos (AR)

Estos modelos aventajan a los anteriores a la hora de reflejar en su formulación tanto la evolución temporal de las características como la correlación entre las componentes. El modelo DAR asume características independientes y es un caso particular del modelo MAR; MAR modela las relaciones entre componentes y DAR en su formulación anula este aspecto.

Las dos subsecciones siguientes quedan dedicadas a la explicación de ambos modelos; aunque se invierte el orden de exposición de creciente dificultad establecido inicialmente, primero se expondrá el modelo MAR de forma detallada para a continuación indicar qué particularizaciones en éste conducen a la formulación del modelo DAR.

2.4.2.4.1. Modelo MAR

El modelo MAR es un modelo autorregresivo multivariante que no asume la independencia entre características, por tanto, y como ya se ha referido en la introducción previa, modela la relación entre ellas. El modelado de las características se plasma en la introducción de matrices donde quedan reflejadas las covarianzas cruzadas de los elementos.

Inicialmente, MAR modela un vector de características estacionario [Meng et al., 2007], de la forma

$$\mathbf{c}_n = \sum_{q=1}^{Q} \mathbf{A}_q \, \mathbf{c}_{n-I(q)} + \mathbf{u}_n \qquad q = 1,\ldots,Q; \qquad (2.13)$$

donde $\mathbf{u}_n$ es un término de ruido independiente e idénticamente distribuido caracterizado por el vector de medias $\mathbf{m}$ y la matriz de covarianzas **C;** *I* es el conjunto de vectores a partir del cual se estimará $\mathbf{c}_n$. Por último, $\mathbf{A}_q$ es la matriz de coeficientes autorregresivos de orden *q*, en la que se encuentran cuantificadas las relaciones entre los distintos componentes. $\mathbf{c}_n$ es el n-ésimo coeficiente MFCC estimado a partir de los *I MFCC* anteriores.

La predicción de $\mathbf{A}_q$ busca minimizar el promedio de los residuos no sólo para un MFCC aislado, sino para todos los MFCC de la ventana de mayor escala temporal. Para cada ventana de mayor duración se obtiene un único vector de parámetros, los parámetros del modelo AR.

Los parámetros de este modelo se obtendrán mediante mínimos cuadrados, con cuidado de introducir un término en el que queden reflejadas las covarianzas cruzadas. El resultado será la consecución de las matrices de coeficientes $\{\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_Q\}$, **m** y **C**, donde $\{\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_Q\} = \arg\min \sum \|u_n\|^2$.

Por último, el vector de integración temporal de ventana *k* para el modelo MAR queda expresado como sigue:

$$\mathbf{z}_k = \left[vec(\mathbf{B}_k)^T, \mathbf{m}_k^T triag(V_k)^T\right]^T, \quad \mathbf{B} = \left[\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_P\right] \tag{2.14}$$

$triag(\cdot)$ indica que, de la matriz de covarianzas de los datos enventanados por la ventana *k* $\boldsymbol{V}_k$ tan solo se toma el triángulo superior o inferior, pero siempre con la diagonal incluida.

2.4.2.4.2. Modelo DAR

Como se ha visto en la subsección 2.4.2.4.1., el modelo MAR asume la interrelación entre características, por lo que en el desarrollo se trata con matrices que modelan dichas interrelaciones. El modelo DAR es una particularización del modelo MAR en el que se asume independencia entre características, que implica covarianzas nulas; el resultado serán unas matrices de coeficientes autorregresivos y de covarianzas diagonales.

En el modelo DAR de orden P los datos enventanados se reescriben como el sumatorio

$$x_n = \sum_{p=1}^{P} a_p x_{n-p} + Gu_n, \quad p = 1, \ldots, P \qquad n = 0, \ldots, W\text{-}1 \tag{2.15}$$

donde a_p son los coeficientes del modelo, el término u_n ruido independiente e idénticamente distribuido de varianza unidad y media *m* y G un factor de escala de u_n.
W es el tamaño de la ventana expresado en número de muestras. Como puede observarse, x_n se modela como una función lineal de salidas previas sumadas a términos de ruido. Nótese también que, a diferencia del modelo MAR, la formulación ya no se fundamenta en vectores, sino en escalares.
Este modelo se ajustará para cada MFCC.

Para estimar los parámetros del modelo DAR, presuponiendo ruido blanco, se volverá a hacer uso del método de mínimos cuadrados y posterior minimización del error, aunque existen otros métodos aplicables tanto en el dominio frecuencial como en el temporal.

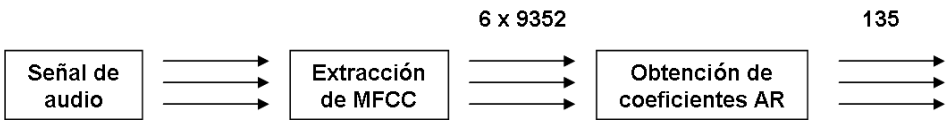
La propiedad extraída de [Meng et al., 2007] y de [Meng, 2006]**,** es la siguiente: La estimación de la señal realizada con los parámetros del modelo autorregresivo es equivalente al espectro real de los vectores de MFCCs, pero suavizado.

2.4.3. Conclusiones

De los métodos de integración temporal presentados y dadas las prestaciones que ofrece, demostradas en [Meng et al., 2007] y [Meng, 2006], en la fase de integración temporal de este Proyecto se ha empleado el modelo MAR.

Una de las ventajas expuestas de la integración temporal es la compactación de los datos: Tomando como ejemplo cifras de archivos de características utilizados en este Proyecto y como puede verse en la Figura 15, la extracción de MFCC dio como resultado una matriz de dimensiones aproximadas de 13×9352. Como ya se ha visto en el apartado 2.2. de este mismo Capítulo es suficiente el empleo de seis MFCC por cada archivo, por tanto en los experimentos la matriz útil de MFCC es de dimensiones 6×9352. Tras realizarse la integración con distintas ventanas temporales, con duraciones comprendidas entre 1000 y 30000 ms. y obteniéndose en todos los casos 135 coeficientes AR, las matrices de coeficientes para una ventana de 1000 ms. tienen unas dimensiones aproximadas de 135×142, mientras que con la ventana temporal de máximo tamaño las matrices se reducen a aproximadamente 135×3. En cualquier caso, incluso en el menos favorable - ventana temporal menor -, la compactación de datos es notable.

Figura 15. Componentes de las etapas de extracción



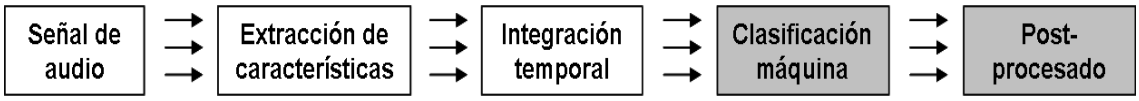
Número de componentes en cada una de las etapas de extracción de coeficientes.

CAPÍTULO 3. Clasificación Máquina

3.1. Introducción

En el Capítulo 1 fueron expuestas de forma muy sucinta nociones relativas a la clasificación máquina, última etapa que se tratará de forma conceptual en esta memoria.

Figura 16. Proceso de clasificación general: Clasificación máquina y post-procesado



Situación de la clasificación máquina, objeto del Capítulo 3, dentro del proceso de reconocimiento automático.

El presente capítulo tiene por objetivo desarrollar los conceptos que se han manejado en la implementación del Proyecto, dividiéndose en los siguientes apartados: Partiendo de la visión general del aprendizaje supervisado presentada en el Capítulo 1, se mostrarán las formulaciones de los métodos de núcleos, cuya cualidad de derivar algoritmos no lineales a partir de otros lineales ha permitido hacer un amplio uso de ellos. A continuación se tratará la versión núcleo reducida del Método de Mínimos Cuadrados Parciales Ortonormalizados (reduced Kernel Orthonormalized Least Squares, rKOPLS en adelante), empleada en los experimentos y derivada de la versión núcleo del Método de Mínimos Cuadrados Parciales Ortonormalizados (Kernel Orthonormalized Least Squares, KOPLS en adelante), analizada también en el último apartado. Por último, se introducirá el post-procesado, que se tratará con mayor precisión en el capítulo siguiente.

3.2. Métodos de Núcleos

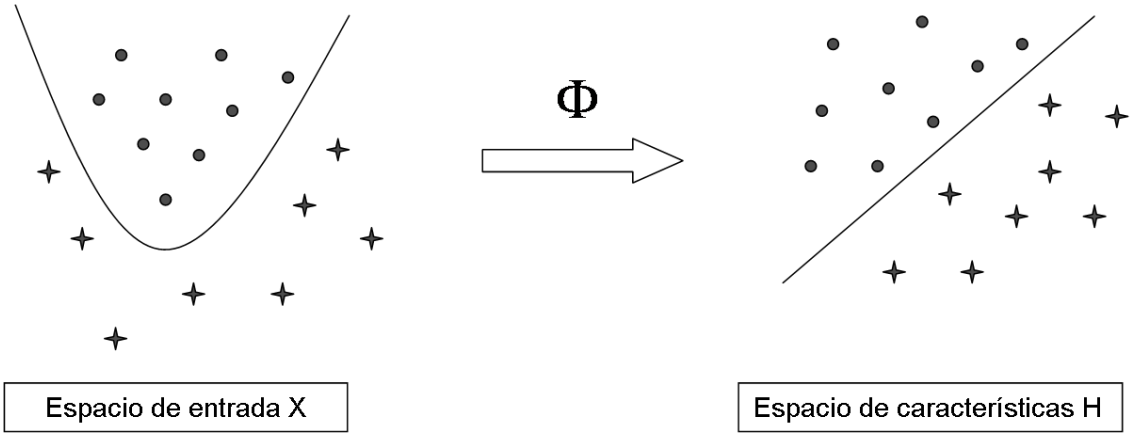
3.2.1. Introducción a los Métodos de Núcleos

Los métodos de núcleos son un conjunto de algoritmos muy empleados en los ámbitos del aprendizaje máquina y del reconocimiento de patrones, cuya principal cualidad es que posibilitan la obtención de algoritmos no lineales a partir de aquéllos lineales mediante una transformación a un espacio de características [Bousquet y Pérez-Cruz, 2003], aumentando de forma significativa la capacidad expresiva de las versiones lineales.

Para conseguir la condición de no linealidad se mapean los datos del algoritmo lineal existentes en un espacio $\mathcal{X}$ a un espacio vectorial $\mathcal{H}$, el espacio de características, mediante una transformación no lineal $\Phi: \mathcal{X} \rightarrow \mathcal{H}$ [Bousquet y Pérez-Cruz, 2003], [Schölkopf et al., 1999] y [Schölkopf y Smola, 2002]. Posteriormente, se ejecuta el algoritmo lineal con la representación $\Phi(x)$, es decir, se efectúa un análisis no lineal de los datos utilizando un método lineal en $\mathcal{H}$.

El propósito del mapeo en Φ es transformar estructuras no lineales de datos en lineales dentro del espacio $\mathcal{H}$, como puede observarse en el ejemplo de la Figura 17.

Figura 17. Transformación según los métodos de núcleos



Transformación de la estructura no lineal según la formulación de los métodos de núcleos.

Hallar la función Φ adecuada puede implicar utilizar un elevado número de dimensiones, incluso llegar a infinitas; es en este punto donde los métodos de núcleos aportan sus cualidades, pues permiten el uso de espacios de características con un número de dimensiones que crece de forma exponencial con el número de variables o incluso infinito [Shawe-Taylor y Cristianini, 2004]. Por otra parte, es necesario destacar que existe un compromiso con el número de dimensiones, dado que si bien con dimensión infinita se puede clasificar linealmente cualquier conjunto de datos, el hecho de disponer de un número excesivo puede llevar al sobreajuste de los datos de entrenamiento y a un extraordinario aumento de la complejidad. Alternativamente, el problema de la dimensionalidad puede atacarse utilizando técnicas de regularización [Bishop, 1995].

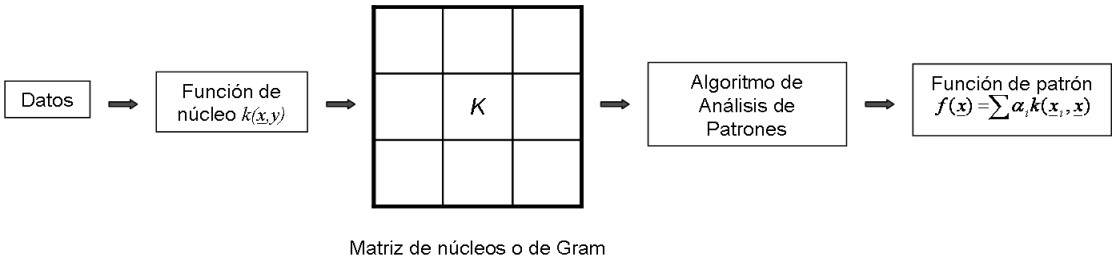
La función de núcleo k corresponde al producto escalar en $\mathcal{H}$, y por tanto se define como:

$$k(\mathbf{x},\mathbf{y}) = (\Phi(\mathbf{x}) \cdot \Phi(\mathbf{y})). \qquad (3.1)$$

Con una reformulación adecuada del algoritmo lineal, la evaluación de la correspondiente función de decisión tan sólo se requiere la evaluación de productos escalares $(\Phi(\mathbf{x})\cdot\Phi(\mathbf{y}))$, y no los patrones mapeados $\Phi(\mathbf{x})$ de forma explícita; este procedimiento se conoce como 'truco del núcleo'. Además, para posibilitar la formulación kernel de un algoritmo lineal es también necesario que la fase de aprendizaje puede ser expresada de forma tal que únicamente se requiere evaluar productos escalares en $\mathcal{H}$ (i.e. formulación kernel).

La correspondencia de los núcleos con los productos internos en un espacio hace que puedan ser considerados como medidas de similitud entre dos puntos. Los productos escalares de pares de datos de entrenamiento evaluados con una función de núcleo se agrupan en la matriz de Gram, a la que será necesario remitirse en caso de demandar cualquier información concerniente al algoritmo de núcleo empleado, véase la distribución de los datos o el nivel de ruido. [Shawe-Taylor y Cristianini, 2004]. De esta forma, el proceso quedaría descrito así:

Figura 18. Proceso de aplicación de los métodos de núcleo



Proceso de aplicación de los métodos de núcleo. A los vectores de parámetros de audio - e.g. vectores AR - se les aplica una función de núcleo que consiga transformarlos de forma adecuada para la posterior decisión. Los productos internos fruto de este cálculo se almacenan ordenadamente en la matriz de Gram, donde serán sometidos a un análisis de búsqueda de patrones, de forma que, a partir de ellos, pueda implementarse una función mediante la cual puedan ser reconocidos y clasificados ejemplos no utilizados durante el entrenamiento. Fuente: [Shawe-Taylor y Cristianini, 2004].

3.2.2. Tipos de núcleos

Antes de enumerar los tipos de núcleos más comunes se debe identificar qué funciones son núcleos. Los núcleos deben cumplir el Teorema de Mercer. Algunos de los núcleos más frecuentemente utilizados, además del lineal $k(\mathbf{x},\mathbf{y}) = (\mathbf{x}\cdot\mathbf{y}) = \mathbf{x}^T\mathbf{y}$, se describen a continuación:

a) Núcleo polinómico

El núcleo polinómico tiene la forma

$$k(\mathbf{x},\mathbf{y}) = (\mathbf{x}\cdot\mathbf{y})^d \tag{3.2}$$

y muestra la correspondencia de un mapeado Φ en un espacio que se extiende por todos los productos de exactamente orden *d* en $\Re^N$.

Este núcleo posee variantes más sofisticadas, por ejemplo

$$k(\mathbf{x},\mathbf{y}) = (\mathbf{x}\cdot\mathbf{y}+c)^d \text{ , con } c > 0, \tag{3.3}$$

con el que se tienen en cuenta todos los productos cuyo orden sea menor o igual que *d*.

b) Núcleo gaussiano

El núcleo gaussiano se define como

$$k(\mathbf{x},\mathbf{y}) = \exp\left(\frac{\|\mathbf{x}-\mathbf{y}\|^2}{2\sigma^2}\right) \qquad \sigma > 0 \tag{3.4}$$

El parámetro σ [Shawe-Taylor y Cristianini, 2004] controla la anchura del núcleo. Cuanto mayor sea el valor de σ mayor será la distancia a la que la exponencial se hace cero. Dado que el kernel tiene un sentido de "similitud", se puede concluir que mientras menor es el valor de sigma más cercanas tienen que estar las muestras para que se las considere "similares".

c) Composición de otros núcleos

Los distintos tipos de núcleos pueden combinarse entre sí para aprovechar al máximo las características de cada uno. Siempre que las operaciones garanticen la verificación del teorema de Mercer, el resultado será un núcleo definido en un nuevo espacio de Hilbert; un ejemplo es el mismo núcleo gaussiano. Una exposición extensa de esta propiedad de los núcleos puede encontrarse en [Shawe-Taylor y Cristianini, 2004].

3.3. rKOPLS

3.3.1. Visión General de los algoritmos de Mínimos Cuadrados Parciales

La dimensión de los datos implicados en los procesos de entrenamiento y test es un parámetro clave, dado que una dimensión elevada aumenta tanto la complejidad del problema como el gasto de recursos. Por añadidura, se asume que la información contenida en los datos sobrepasa la representación necesaria para la posterior clasificación, por tanto la dimensión de éstos debe ser reducida. Es en este punto donde los Mínimos Cuadrados Parciales ('Partial Least Squares', PLS en adelante) aportan sus cualidades de análisis de relaciones entre datos, haciendo uso de variables latentes, [Arenas-García et al., 2006] [Nielsen et al., 2007] es decir, de variables ocultas que, en el caso concreto de datos de audio etiquetados, proporcionan información tanto de los datos como de las etiquetas. Los PLS se caracterizan por su capacidad para tratar de forma robusta conjuntos de datos cuyo número de dimensiones excede al de muestras.

El algoritmo PLS básico considera dos conjuntos de datos **X** e **Y**, características de los datos de audio y etiquetas respectivamente, en los que las muestras se disponen de forma ordenada, en filas o en columnas; mediante un procedimiento iterativo o bien con un planteamiento basado en autovalores se trata de encontrar variables latentes que den cuenta de la covarianza entre ambos conjuntos de datos. Los conjuntos **X** e **Y** se transforman mediante un proceso que sustrae la información contenida en las variables latentes.

En los primeros enfoques de los PLS la relación entre las variables latentes de **X** y de **Y** debía ser lineal, lo que limitaba fuertemente su aplicación; sin embargo, con la aparición de los métodos de núcleos se puede obtener la flexibilidad de expresiones no lineales al tiempo que se manejan tan sólo ecuaciones lineales. En el campo de los PLS se han producido reformulaciones para dotarlos de potencialidades núcleo [Rosipal y Trejo, 2001]**,** [Arenas-García et al., 2006]. De manera análoga a la formulación vista anteriormente, los datos de entrada se mapean utilizando una función no lineal en un espacio de alta dimensión en el cual el PLS lineal actúa sobre los datos transformados. Por otro lado, se explota la propiedad del 'truco del núcleo' (cf. (3.1)).

A pesar de las bondades de los PLS, son métodos a los que se puede imputar ciertas desventajas. Una de ellas, común a otros métodos basados en núcleos, es la dimensión de las matrices, pues para un conjunto de l muestras, las matrices núcleo utilizadas serán $l \times l$, por tanto ha de considerarse el número de muestras que se incluyen en los experimentos para evitar problemas de memoria y tiempo de ejecución, tanto en la fase de entrenamiento como de test [Arenas-García et al., 2006] [Nielsen, 2007].

La técnica utilizada en este Proyecto ha sido la de rKOPLS (*reduced Kernel Orthonormalized Partial Least Squares*), variante de los PLS con propiedades núcleo especialmente útil en la extracción de características en bases de datos de alta dimensión. Consta de dos partes: Primero, una variante ortonormalizada de los PLS con características núcleo llamada KOPLS, en la que se transforman los datos de entrada de manera que se mantengan ortonormales entre sí, del mismo modo que las proyecciones; segundo, un enfoque disperso para conjuntos de datos de gran tamaño.

3.3.2. KOPLS

Los datos con los que se inicia el desarrollo de KOPLS son un conjunto de pares $\{\phi(\mathbf{x}_i), \mathbf{y}_i\}_{i=1}^{\ell}$, $\mathbf{x}_i \in \Re^N, \mathbf{y}_i \in \Re^M$ y la función $\phi(\mathbf{x}) : \Re^N \rightarrow \mathcal{H}$ que mapea los datos de entrada en un espacio de Hilbert - espacio de características - asociado a un núcleo y cuya dimensión normalmente será muy grande o incluso infinita. A partir de ellos se conforman las matrices $\mathbf{\Phi} = [\phi(\mathbf{x}_1), \ldots, \phi(\mathbf{x}_\ell)]^T$ e $\mathbf{Y} = [\boldsymbol{y}_1, \ldots, \boldsymbol{y}_l]$, donde $\mathbf{\Phi}$ es la matriz que alberga las transformaciones realizadas a los datos de audio en el espacio de características e $\mathbf{Y}$ la matriz de etiquetas. Se desea construir las matrices $\mathbf{\Phi}' = \mathbf{\Phi U}$ e $\mathbf{Y}' = \mathbf{YV}$. Cada una de ellas contiene n_p proyecciones de los datos originales de entrada y salida; en cuanto a $\mathbf{U}$ y $\mathbf{V}$, se definen como las matrices de proyecciones de dimensiones $\dim(\mathcal{H}) \times n_p$ y $M \times n_p$, respectivamente. El objetivo de los algoritmos de análisis multivariante de núcleos es rastrear las matrices de proyecciones de tal forma que los datos de entrada y salida proyectados se alineen lo máximo posible.

En el presente Proyecto se utiliza una extensión núcleo de un método de análisis multivariante denominado Mínimos Cuadrados Parciales Ortonormalizados (Orthonormalized Partial Least Squares, OPLS). La variante núcleo, KOPLS (Kernel Orthonormalized Partial Least Squares, KOPLS), se puede establecer desde dos planteamientos equivalentes. Se partirá del más intuitivo conceptualmente para luego plantear un problema de maximización cuyo objetivo será encontrar el mayor parecido entre los datos proyectados y las etiquetas.

La formulación más intuitiva se apoya en que KOPLS extrae proyecciones de los datos de entrada y proporciona proyecciones óptimas para la multirregresión lineal efectuada en el espacio de características, por lo que se buscará una solución que minimice la suma de cuadrados de los residuos de la aproximación a la matriz de etiquetas. Este planteamiento puede expresarse como sigue:

$$\left\| \tilde{\mathbf{Y}} - \tilde{\mathbf{\Phi}}' \hat{\mathbf{B}} \right\|_F^2, \qquad \text{con } \hat{\mathbf{B}} = (\tilde{\mathbf{\Phi}}'^{\mathbf{T}} \tilde{\mathbf{\Phi}}')^{-1} \tilde{\mathbf{\Phi}}'^{T} \tilde{\mathbf{Y}}, \tag{3.5}$$

donde $\| \cdot \|_F$ denota la norma de Frobenius de una matriz y $\hat{\mathbf{B}}$ es la matriz de regresión óptima. $\tilde{\mathbf{\Phi}}$ e $\tilde{\mathbf{Y}}$ son versiones centradas de $\mathbf{\Phi}$ e $\mathbf{Y}$, (es decir, a las que se les ha restado la media de cada columna) respectivamente. El hecho de centrar las matrices $\mathbf{\Phi}$ e $\mathbf{Y}$ supone trasladar el origen del espacio de características al centro de masa de los datos.

Por ende, centrar las matrices $\mathbf{\Phi}$ e $\mathbf{Y}$ minimiza la suma de autovalores de ambas matrices[2] [Shawe-Taylor y Cristianini, 2004]. El superíndice T denota la transposición del vector o de la matriz.

El objetivo es minimizar la función de coste expuesta con respecto de la matriz de proyecciones **U**.

La formulación a partir de la cual se planteará el problema de maximización se expondrá en las siguientes líneas, sin embargo la equivalencia entre ambas no se demostrará analíticamente en este Proyecto. Si se desea consultar la demostración desarrollada puede encontrarse en [Shawe-Taylor y Cristianini, 2004].

Como ya se ha mencionado anteriormente, el objetivo último es encontrar el máximo parecido entre los datos proyectados $\mathbf{U^T\Phi}$ y la matriz de etiquetas **Y**. Para ello, se plantea el siguiente problema de maximización:

KOPLS: Maximizar $Tr\{\mathbf{U^T\tilde{\Phi}^T\tilde{Y}\tilde{Y}^T\tilde{\Phi}U}\}$

Con la restricción $\mathbf{U^T\tilde{\Phi}^T\tilde{\Phi}U} = \mathbf{I}$ (3.6)

Al igual que en el planteamiento anterior, $\tilde{\mathbf{\Phi}}$ e $\tilde{\mathbf{Y}}$ son versiones centradas de $\mathbf{\Phi}$ e **Y**, **I** es la matriz identidad de dimensión n_p y el superíndice T denota la transposición del vector o la matriz. Mediante la restricción se asegura que las distintas proyecciones de los datos estarán incorrelacionadas entre sí.

De nuevo el objetivo será maximizar la función de coste con respecto de **U**.

A la vista de ambos planteamientos puede apreciarse que KOPLS no sólo es útil para problemas de multirregresión, sino que también constituye un extractor de características en versión núcleo de gran potencia.

El problema de maximización planteado en (3.6) puede reescribirse si se hace uso del Teorema de Representación [Haykin, 1999], según el cual los vectores columna de **U** (matriz de proyecciones cuya dimensión es dim($\mathcal{H}$) $\times n_p$) pueden expresarse como combinación lineal de los datos de entrenamiento. Si se sustituye **U** por $\tilde{\mathbf{\Phi}}^T$ A en (3.6), donde A = $[\boldsymbol{\alpha}_1,\ldots,\boldsymbol{\alpha}_{n_p}]$ y $\boldsymbol{\alpha}_i$ es un vector columna de longitud *l* que contiene los coeficientes para el vector de proyección i-ésimo, el problema quedará formulado así:

Maximizar $Tr\{\mathbf{A^T K_x K_y K_x A}\}$

con la restricción $\mathbf{A^T K_x K_x A} = \mathbf{I}$ (3.7)

En este caso la traza deberá maximizarse con respecto de **A.**

[2] El centro de masa es el punto en el cual la suma de las normas de los puntos del espacio es mínima; la suma de las normas es la traza de la matriz, por tanto es equivalente a la suma de los autovalores.

Las matrices $\mathbf{K}_x$ y $\mathbf{K}_y$ se definen como $\tilde{\mathbf{\Phi}}\tilde{\mathbf{\Phi}}^T$ e $\tilde{\mathbf{Y}}\tilde{\mathbf{Y}}^T$ respectivamente. Por el hecho de formularse como núcleos sólo se trabaja con los productos internos del espacio $\mathcal{H}$, y no con los datos, una de las grandes ventajas de los métodos de núcleos. De hecho, aprovechando esta cualidad, esta reescritura es particularmente útil cuando el número de dimensiones de $\phi(\mathbf{x})$ es infinito (recordemos que $\mathbf{\Phi} = [\phi(\mathbf{x}_1),\ldots,\phi(\mathbf{x}_\ell)]^T$), pues permite calcular los elementos de la matriz de kernel incluso cuando cada una de las infinitas dimensiones de $\phi(\mathbf{x})$ no pueden ser expresadas.

Aplicando álgebra lineal a (3.7) las columnas de **A** vendrán como solución del siguiente problema de autovalores generalizados:

$$\mathbf{K}_x\mathbf{K}_y\mathbf{K}_x\boldsymbol{\alpha} = \lambda\mathbf{K}_x\mathbf{K}_x\boldsymbol{\alpha} \tag{3.8}$$

El procedimiento elegido en [Arenas-García et al., 2006] para resolver este problema consiste en calcular iterativamente el vector de proyección óptimo, y posteriormente disminuir el rango de las matrices involucradas. Esta es, de forma resumida, el proceso en el paso *i*-ésimo:

1. Encontrar el mayor autovalor generalizado de (3.8), así como su autovector generalizado correspondiente: $\{\lambda_i, \boldsymbol{\alpha}_\mathbf{i}\}$. $\boldsymbol{\alpha}_\mathbf{i}$ se normalizará para satisfacer la condición $\boldsymbol{\alpha}_\mathbf{i}\mathbf{K}_x\mathbf{K}_x\boldsymbol{\alpha}_\mathbf{i} = 1$.

2. Disminuir el rango de la matriz $l \times l$ $\mathbf{K}_x\mathbf{K}_y\mathbf{K}_x$ de acuerdo a:

$$\mathbf{K}_x\mathbf{K}_y\mathbf{K}_x \leftarrow \mathbf{K}_x\mathbf{K}_y\mathbf{K}_x - \lambda_i\mathbf{K}_x\mathbf{K}_x\boldsymbol{\alpha}_\mathbf{i}\boldsymbol{\alpha}_\mathbf{i}^T\mathbf{K}_x\mathbf{K}_x,$$

   de esta forma se eliminará de la matriz de etiquetas **Y** la mejor aproximación, basada en las proyecciones computadas en el paso i-ésimo; puesto que la información relativa a las variables latentes ya ha sido extraída, los vectores involucrados no aportarán nueva información en los pasos siguientes. Con este esquema el rango de $\mathbf{K}_x\mathbf{K}_y\mathbf{K}_x$ decrece una unidad en cada paso. Dado que el rango de la matriz original $\mathbf{K}_y$ es como máximo el rango de **Y**, éste será el mayor número de proyecciones que se pueden derivar al utilizar KOPLS.

Este algoritmo iterativo, que es muy similar a los utilizados en otros enfoques de análisis multivariante, tiene la ventaja de que en cada iteración se logra la solución óptima con respecto al número de proyecciones extraídas.

3.3.3. Enfoque Compacto de la Solución KOPLS: rKOPLS

La formulación núcleo del algoritmo OPLS presentado adolece de varias limitaciones; como ocurre en otros métodos núcleo, KOPLS requiere la implementación y almacenamiento de una matriz de núcleos $l \times l$, que limita de forma drástica el tamaño de los conjuntos de datos en los que el algoritmo puede aplicarse. Por otra parte, los procedimientos algebraicos empleados para resolver el problema de autovalores generalizados (cf. 3.8) necesitan invertir la matriz $\mathbf{K_x K_x}$, que puede no alcanzar el rango máximo posible, imposibilitando así dicha inversión. Por último, la matriz **A** generalmente no será dispersa, lo que supone calcular los núcleos entre los nuevos datos y todas las muestras del conjunto de entrenamiento cuando los datos deban proyectarse, proceso de tremenda ineficiencia tanto en tiempo como en recursos.

El criterio empleado para solventar estos inconvenientes será imponer dispersión en la representación de los vectores de proyección, por ejemplo utilizando la aproximación $\mathbf{U} = \mathbf{\Phi}_\mathbf{R}^\mathbf{T}\mathbf{B}$, donde $\mathbf{\Phi_R}$ es un subconjunto de R (R < l) patrones seleccionados de forma aleatoria procedentes de los datos de entrenamiento y **B** = $[\beta_1,\ldots,\beta_{n_p}]$ contiene los parámetros del modelo compacto. En contraposición al modelo KOPLS, donde **U** = $\tilde{\mathbf{\Phi}}^T\mathbf{A}$ y $\mathbf{\Phi}$ tiene tantas columnas como puntos de entrenamiento, ahora $\mathbf{\Phi_R}$ tiene únicamente R columnas. Como consecuencia inmediata se reducen de forma considerable las dimensiones de las matrices: mientras en el modelo KOPLS $\mathbf{K_x} = \tilde{\mathbf{\Phi}}\tilde{\mathbf{\Phi}}^\mathbf{T}$ la dimensión es de $l \times l$, con esta nueva aproximación la dimensión de $\mathbf{K_R} = \mathbf{\Phi_R}\tilde{\mathbf{\Phi}}^\mathbf{T}$ es de $R \times l$, con lo que se disminuye el número de kernels que se deben calcular y almacenar, concretamente de l^2 a $l \times R$. Nótese la importante reducción de operaciones de cálculo si R es sensiblemente menor que el número de datos. A pesar de estas reducciones tanto de dimensiones como de operaciones, la aproximación puede ser muy similar con respecto de la que pudiera hacerse con todos los datos de entrenamiento. Por otro lado, los datos pueden quedar correctamente caracterizados con un número menor de parámetros.

Si se reescribe (3.6) con la aproximación de $\mathbf{U} = \mathbf{\Phi}_\mathbf{R}^\mathbf{T}\mathbf{B}$ se llegará a un problema de maximización alternativo que constituye la base de rKOPLS:

rKOPLS: Maximizar $Tr\{\mathbf{B}^\mathbf{T}\mathbf{K_R}\mathbf{K_y}\mathbf{K_R}^\mathbf{T}\mathbf{B}\}$, con $\mathbf{K_R} = \mathbf{\Phi_R}\tilde{\mathbf{\Phi}}^\mathbf{T}$

$$\text{Con la restricción } Tr\{\mathbf{B}^\mathbf{T}\mathbf{K_R}\mathbf{K_R}^\mathbf{T}\mathbf{B}\} = \mathbf{I} \qquad (3.9)$$

En esta definición, $\mathbf{K_R}$ es una matriz de núcleos reducida, de dimensiones $R \times l$. En este caso, la maximización se realiza con respecto de **B**.

De forma similar al algoritmo KOPLS, las proyecciones para rKOPLS se pueden obtener mediante la resolución del problema de autovectores generalizados:

$$\mathbf{K_R}\mathbf{K_y}\mathbf{K_R}^\mathbf{T}\boldsymbol{\beta} = \lambda\mathbf{K_R}\mathbf{K_R}^\mathbf{T}\boldsymbol{\beta} \qquad (3.10)$$

El procedimiento iterativo presentado al final del subapartado anterior continúa siendo válido para rKOPLS, efectuando algunas modificaciones:

1. Encontrar el mayor autovalor generalizado de (3.10), así como su autovector generalizado correspondiente: $\{\lambda_i, \boldsymbol{\beta}_i\}$. También aquí será necesario normalizar $\boldsymbol{\beta}_i$ para satisfacer la condición $\boldsymbol{\beta}_i \mathbf{K}_R \mathbf{K}_R^T \boldsymbol{\beta}_i = 1$.
2. Disminuir el rango de la matriz $R \times R$ $K_R K_y K_R^T$ de acuerdo a:

$$\mathbf{K}_R \mathbf{K}_y \mathbf{K}_R^T \leftarrow \mathbf{K}_R \mathbf{K}_y \mathbf{K}_R^T - \lambda_i \mathbf{K}_R \mathbf{K}_R^T \boldsymbol{\beta}_i \boldsymbol{\beta}_i^T \mathbf{K}_R \mathbf{K}_R^T$$

Para terminar, se exponen las propiedades de rKOPLS, mostrando a su vez cómo supera algunas de las limitaciones de KOPLS estándar.

- A diferencia de KOPLS, rKOPLS se basa en una solución dispersa, de manera que los nuevos datos se proyectan con sólo $R < l$ evaluaciones del núcleo por patrón, lo que resulta especialmente ventajoso con conjuntos grandes de datos de entrenamiento.

- Las proyecciones de entrenamiento de rKOPLS sólo necesitan el cálculo de la matriz de núcleo reducido $\mathbf{K}_R$, de dimensiones $R \times l$.

- El algoritmo rKOPLS sólo emplea las matrices $\mathbf{K}_R \mathbf{K}_R^T$ y $\mathbf{K}_R \mathbf{K}_y \mathbf{K}_R^T$. Ambas matrices, de dimensión $R \times R$, se pueden calcular sin recurrir a la implementación explícita de $\mathbf{K}_R$, reduciendo así de forma notoria los requerimientos de memoria; de nuevo una propiedad muy conveniente cuando se trabaja con conjuntos de datos grandes.

- Por último, para determinados valores de σ $\mathbf{K}_R \mathbf{K}_R^T$ es de rango deficiente. En el enfoque de rKOPLS el parámetro R actúa como regularizador, cuando es suficientemente pequeño hace que $\mathbf{K}_R \mathbf{K}_R^T$ alcance el rango máximo y en consecuencia su inversión no plantee dificultades.

A tenor de las ventajas aportadas por rKOPLS, tanto en lo referente a la reducción de la dimensión como en la mayor consistencia del proceso, y de los resultados de los experimentos realizados en [Arenas-García et al., 2006], en este Proyecto se ha optado por incluir esta alternativa de los métodos PLS en la implementación de las funciones conducentes a completar el proceso de clasificación máquina, pues rKOPLS ofrece resultados comparables a KOPLS además de la consiguiente economía de recursos computacionales.

3.4. Post-procesado

Una vez realizado el proceso de clasificación máquina se plantea el problema de a qué clase debe asignarse la canción completa, puesto que para cada canción existen varios vectores de alto nivel, y cada uno de ellos se asignará de acuerdo al artista o al género. De ahí la necesidad del post-procesado.

Un ejemplo de post-procesado consiste en combinar las salidas de varios clasificadores y en función de ellas ofrecer una única. Otra técnica, denominada fusión temporal (*temporal fusion*) y que se aplica en la decisión en escalas temporales mayores, consiste en combinar la secuencia de salidas de un clasificador en una decisión consensuada única.

En este Proyecto se ha utilizado como técnica de post-procesado el voting, que a grandes rasgos consiste en, dada una serie de experimentos con una base de datos dada, construir una matriz con todos los resultados y seleccionar, para cada uno de los archivos que conforma la base de datos, el género o artista que haya obtenido mejor resultado y que por tanto, se supone que es la característica de alto nivel que describe al archivo. Esta técnica de post-procesado se verá en detalle en el capítulo dedicado a los experimentos.

CAPÍTULO 4. Experimentos

4.1. Introducción

Este capítulo está dedicado a la descripción y discusión de los experimentos. El objetivo de esta fase es analizar la influencia del tamaño de la ventana temporal a la hora de representar las características de alto nivel partiendo de las de bajo nivel. Esta idea es el hilo argumental que hilvana todo el Proyecto y que determina decisivamente su estructura, como podrá verse en la descripción que se realizará de los elementos y del proceso conducente a las curvas de validación y test.

El capítulo estará estructurado de la siguiente forma: En primer lugar se ofrecerá una descripción de las bases de datos que se han utilizado. A continuación se relatará el proceso seguido para la consecución de los resultados, con especial énfasis en los ajustes para los selectores de características y clasificadores. Por último, se analizarán los resultados y se extraerán las conclusiones pertinentes.

4.2. Descripción de las bases de datos

A continuación se hará un repaso sobre bases de datos utilizadas en investigaciones anteriores cuyas características o deficiencias han ayudado a definir la forma más idónea de presentar los datos para experimentos de reconocimiento musical. Una vez contextualizados los antecedentes de este campo de trabajo se describirán las bases de datos utilizadas en este Proyecto.

4.2.1. Parámetros de los archivos de audio

4.2.1.1. Descripción y Parámetros de la Base de Datos de Artistas

Una herramienta esencial en la tarea de clasificar archivos de audio es, por supuesto, la base de datos. Sin embargo, y a pesar del extenso trabajo de investigación que se viene realizando en el campo del reconocimiento y clasificación musical, no existe una gran variedad de conjuntos de datos, si bien es cierto que gracias a la competición anual MIREX (*Music Information Retrieval Evaluation eXchange*) esta situación ha ido variando significativamente.

Uno de los problemas fundamentales en el campo de la recuperación de datos musicales son los impedimentos legales en lo referente a la compartición de archivos musicales. Esto supone en la mayoría de los casos llegar a soluciones subóptimas, pues el acceso limitado a contenidos puede traer como consecuencia bases de datos insuficientemente representadas, carentes de variedad, y por tanto, no válidas como método de evaluación de algoritmos de reconocimiento musical.

Para intentar paliar esta falta de material, el Laboratorio LabROSA del Departamento de Ingeniería Eléctrica de la Universidad de Columbia, encabezado por Daniel P. W. Ellis, ha diseñado un conjunto de datos y tareas que puede descargarse y ejecutarse en cualquier máquina. Esta base de datos es *artist20*.

artist20 es una base de datos formada por seis álbumes de veinte artistas diferentes. El total de pistas de los 120 álbumes es de 1.403. Esta herramienta mejora a la versión anterior, *uspop2002*, donde estaban identificados 18 artistas con cinco o más álbumes. El conjunto de datos USPOP2002 consta de aproximadamente 8700 canciones diferentes distribuidas entre 400 músicos pop norteamericanos, como fruto del trabajo conjunto de LabROSA (Columbia), MIT y laboratorios HP (Cambridge). Estos datos fueron utilizados en la etapa de entrenamiento de la identificación de artistas. No obstante, y a pesar de utilizar álbumes diferentes para entrenamiento y test, esta base de datos presentaba deficiencias como pistas repetidas, grabaciones en vivo, etc.

artist20 se implementó para resolver estos inconvenientes. La mayor parte de los elementos de esta nueva base de datos se toman del mismo *uspop2002*, pero incrementada con varios artistas y álbumes que no se encontraban en ésta última, todo con el fin de obtener seis álbumes por cada artista. También se ha hecho hincapié en evitar cambios de estilo bruscos. Toda la música que compone *artist20* ha sido recogida de las colecciones personales de los miembros del laboratorio y amigos. El nuevo conjunto de datos para clasificación musical consiguió saltar todos los obstáculos legales existentes mediante la inclusión de tan sólo 10 segundos de fragmentos musicales, muestreados de forma aleatoria dentro de cada canción completa.

Como añadido al material descrito, también se distribuyen listas de archivos compuestos de varios cortes de la base, así como un conjunto de entrenamiento canónico (tres álbumes por artista), un conjunto de validación (un álbum) y un conjunto de test (dos álbumes). Por último, se ha definido un corte de seis carpetas para cumplir el esquema entrenamiento/test, en el que cada carpeta está compuesta de cinco álbumes por artista para la etapa de entrenamiento, quedando el restante para la etapa de test. El resultado final viene de promediar todas las carpetas.

La herramienta incluye código Matlab para evaluar las seis carpetas de entrenamiento y test. En términos de codificación, los datos se presentan como MFCCs precalculados (20 MFCCs por archivo con un solapamiento de 10 ms.) y matrices beat-chroma[3] (formadas por 12 vectores chroma), pero también como MP3 mono de 32 kbps

[3] Las características chroma consisten en un vector de doce elementos en el que cada dimensión representa la intensidad asociada a un semitono en particular, con independencia de la octava; de esta forma, se intenta representar la melodía y la armonía mientras se minimiza la influencia de la instrumentación. [Ellis, 2007].

(16 kHz de tasa de muestreo y limitado en banda a 7.2 kHz). Los MFCC y características chroma que se han utilizado arrojan resultados similares a los ofrecidos con datos estéreo a 44 kHz. Por otra parte, la calidad de este formato es comparable a radio AM.

4.2.1.2. Descripción y Parámetros de la Base de Datos de Género

La base de datos de género consta de 1317 elementos distribuidos entre los once géneros. Asimismo, las canciones están extraídas de los repertorios de 720 artistas diferentes, lo que supone aproximadamente una media de dos piezas musicales por artista.

Las piezas musicales están codificadas en formato MPEG1 de capa 3 (estéreo) a una tasa de 128 kbps. En los experimentos, fueron submuestreadas a 22050 Hz

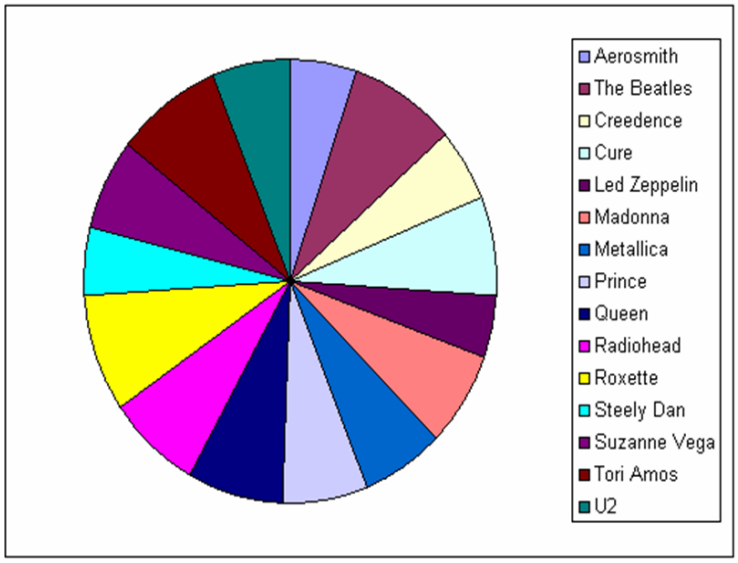
4.2.1.3. Bases de datos en el Proyecto

Los experimentos de este Proyecto se han realizado con la ayuda de dos bases de datos, con las que se trabajará en la clasificación de las características de alto nivel de "Artista" y "Género". La base de datos de artistas consta de 1025 elementos repartidos entre 15 artistas, mientras que la de género está formada por 1309 elementos repartidos entre 11 géneros musicales. Cada uno de los elementos de la base de datos está formado por una matriz de datos, que contiene los coeficientes MFCC del fragmento musical en cuestión, y una etiqueta, que indica a qué género o artista pertenece dicha matriz. La matriz de datos está formada por 13 MFCC, de los cuales en la fase de experimentos se utilizarán 6, siguiendo las conclusiones en [Meng et al., 2005], que indican que con este número los datos quedan suficientemente representados. La etiqueta será un número natural comprendido entre el 1 y el 15 en caso de los artistas y entre el 1 y el 11 en la base de datos de género. Obviamente, el número de etiquetas coincide con el número de clases en cada una de la base de datos.

En ambas bases de datos se busca que los elementos estén distribuidos de forma medianamente uniforme; si se utilizaran bases de datos en los que los elementos se distribuyeran de forma desigual sería imprescindible aplicar algún mecanismo de corrección. Este imperativo se debe a que, por motivos de implementación, es necesario que todas las etiquetas, es decir, todos los géneros y artistas, se encuentren representadas tanto en los datos de entrenamiento como en los de test. Por todo ello, la distribución de los datos es equiprobable, como puede verse en la Figura 19.

Figura 19. Distribución de la base de datos de artistas

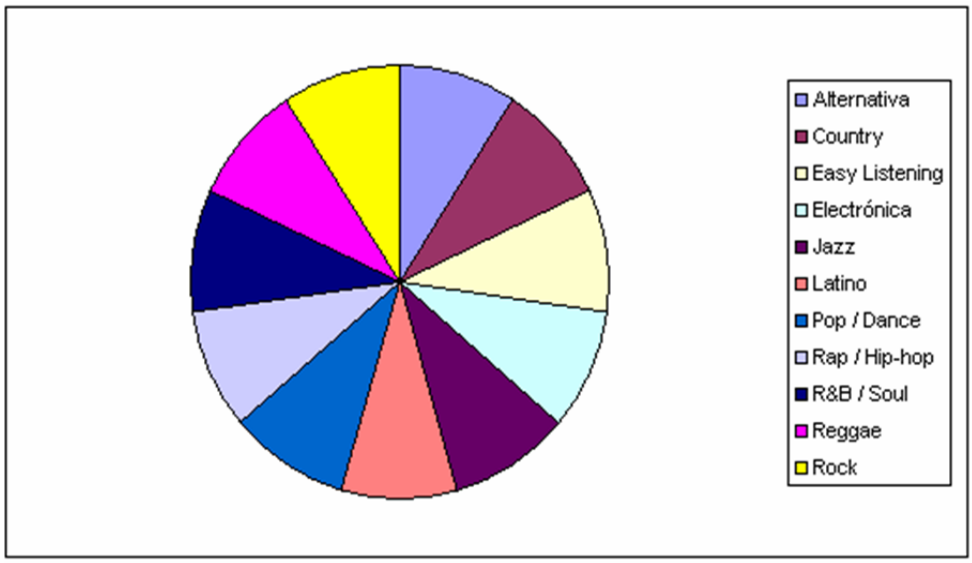
Artista	Número de ítems	Porcentaje en BD
Aerosmith	53	5,2%
The Beatles	86	8,4%
Creedence	55	5,4%
Cure	75	7,3%
Led Zeppelin	44	4,3%
Madonna	71	6,9%
Metallica	65	6,3%
Prince	68	6,6%
Queen	78	7,6%
Radiohead	73	7,1%
Roxette	88	8,6%
Steely Dan	52	5,1%
Suzanne Vega	69	6,7%
Tori Amos	84	8,2%
U2	64	6,2%
Total	1025	100,0%



Distribución numérica y gráfica de los elementos de la base de datos de artistas. Como puede observarse, los 15 artistas se encuentran distribuidos de manera casi uniforme, cumpliéndose así una de las exigencias de la implementación del Proyecto.

Figura 20. Distribución de la base de datos de géneros

Género	Número de ítems	Porcentaje en BD
Alternativa	119	9,1%
Country	119	9,1%
Easy Listening	119	9,1%
Electrónica	119	9,1%
Jazz	119	9,1%
Latino	119	9,1%
Pop / Dance	119	9,1%
Rap / Hip-hop	119	9,1%
R&B / Soul	119	9,1%
Reggae	119	9,1%
Rock	119	9,1%
Total	1309	100,0%



Distribución numérica y gráfica de los elementos de la base de datos de género. Al igual que en la base da datos de artistas, los 11 géneros se encuentran distribuidos de forma cercana a equiprobable, cumpliéndose así una de las exigencias de la implementación del Proyecto.

4.3. Descripción de los experimentos

4.3.1. Consideraciones previas

Antes de proceder a la exposición de los experimentos se van a aclarar algunos parámetros y conceptos generales.

- Las proporciones de datos de entrenamiento y de test utilizadas durante los experimentos son del 80% y 20%, respectivamente. Estos datos se reordenan de forma aleatoria antes de comenzar los experimentos. La partición se hace por canciones, y no por vectores de bajo nivel, para evitar una excesiva redundancia.

- Las ventanas temporales con las que se han llevado a cabo los experimentos se encuentran en el intervalo que va desde los 1.000 ms. hasta los 60.000 ms., con un paso entre ellas de 5.000 ms.

- Parámetro σ: Como ya se expuso en el Subapartado 3.1.2. del Capítulo 3, el parámetro σ controla la flexibilidad del núcleo gaussiano. Este parámetro va en función de la dimensión de los coeficientes AR. Los valores con los que se han realizado los experimentos en este Proyecto van de √dim/√32 a √dim*√1024, con un paso de una potencia de 2 y donde *dim* es la dimensión de los coeficientes AR.

- Validación cruzada: Como ya se ha venido mostrando a lo largo de este Proyecto, uno de los objetivos del aprendizaje de la máquina es minimizar la función de error definida respecto a los datos de entrenamiento. El correcto desempeño de la máquina tras este entrenamiento se verifica mediante la evaluación de la función de error usando un conjunto de datos de validación, independiente del conjunto de datos de entrenamiento. Por último, la capacidad expresiva de la máquina vuelve a confirmarse con un tercer grupo de datos, los datos de test. Sin embargo, en la práctica la disponibilidad de datos etiquetados para el entrenamiento puede ser francamente limitada, con lo que puede ocurrir que no haya una cantidad de datos suficiente como para limitar su uso a una sola de las etapas descritas anteriormente. Es en este escenario donde se recurre a la validación cruzada, que consiste en dividir el conjunto de datos de entrenamiento de forma aleatoria en S subconjuntos de tamaño similar. La fase de entrenamiento se realizará entonces con los datos de S-1 subconjuntos y la fase de test con el conjunto restante. Este proceso se repetirá de forma iterativa S veces, utilizando todas las combinaciones de S-1 subconjuntos como datos de entrenamiento y cada uno de los subconjuntos restantes de la partición anterior como datos de test. La validación cruzada se utiliza para ajustar el valor de los parámetros. Una vez seleccionados se vuelve a ejecutar el entrenamiento y posteriormente se verifican las prestaciones sobre el conjunto de test. Como desventaja puede presentarse el hecho de repetir la iteración S veces, que en caso de un valor relativamente elevado se traduce en consumo de tiempo y recursos computacionales. Información más extensa relativa a la validación cruzada puede encontrarse en [Bishop, 1995].

- Clasificación y esquemas de combinación: Como ya se adelantó de forma sucinta en el último apartado del Capítulo 3, en este Proyecto se han empleado técnicas de post-procesado con el fin de obtener un mayor rendimiento de las capacidades de la máquina. Para cada canción existen varios vectores de alto nivel, y cada uno de ellos se asignará de acuerdo al artista y al género. Ante el problema de a qué clase debe asignarse la canción completa se aporta como solución la etapa de post-procesado.
- Este post-procesado se engloba dentro de los métodos de fusión de información, cuyo propósito es obtener una clasificación de los datos eficiente. Dentro de estos métodos se pueden encontrar los métodos de fusión de información temprana (*early information fusion*), cuyo ejemplo dentro de este Proyecto son los coeficientes AR, y tratan principalmente de modelar las características temporales en o antes del modelo de clasificación estadístico, y los métodos de fusión de información tardía (*late information fusion*), que es el método de combinar los resultados proporcionados por el clasificador. Existen varios esquemas de combinación, como el voto por mayoría, la suma o la media. En el voto por mayoría se cuentan los votos recibidos por el clasificador y la opción con mayor número de ellos es la que se selecciona como resultado, tratándose así de una decisión consensuada. Con la suma las probabilidades asociadas a cada ejemplo se suman y la decisión se basa en ese resultado. La media es un resultado similar a la suma pero tomándose la media en vez de la suma. Para este Proyecto se ha escogido como técnica de post-procesado el voting (voto por mayoría), dado que ninguno de los métodos ha demostrado un mejor desempeño con respecto a los otros en investigaciones de reconocimiento musical y en principio no debería afectar de forma sustancial a las conclusiones de este Proyecto. En [Meng et al., 2005] pueden encontrarse experimentos en los que se comparan todas estas técnicas.

- Cada serie de experimentos se realiza con ambas bases de datos. Asimismo, la partición aleatoria entre elementos de entrenamiento y elementos de test permanece constante durante toda la serie de experimentos.

4.3.2. Algoritmo

La secuencia de pasos que se ha seguido para una duración dada ha sido el siguiente:

a) Selección de la base de datos.
b) Definición de las ventanas temporales, tal y como ha quedado descrito anteriormente.
c) Establecimiento de la dimensión de la matriz de coeficientes AR en función de la duración de la ventana temporal anteriormente citada.
d) Definición de los valores de σ.
e) Mediante el procedimiento de validación cruzada descrito anteriormente se procede al entrenamiento de la máquina. Para ello, se reinicia el extractor de características y se asignan para entrenamiento todos los subconjuntos de datos reordenados de forma aleatoria. Posteriormente se eliminará el de test, distinto en cada iteración.

f) Se identifican las canciones que forman parte de los subconjuntos de datos de entrenamiento y test.
g) Generación de las matrices de AR y etiquetas partiendo de los subconjuntos de datos de entrenamiento y test.
h) De forma iterativa, para todos los valores de σ, se extraen las características de los coeficientes AR de las canciones mediante la utilización de rKOPLS.
i) Se extraen las proyecciones de la matriz de canciones, es decir, la matriz de características. Cada fila será un patrón en el nuevo espacio.
j) Se obtienen los coeficientes de regresión partiendo de la matriz de características y de la matriz de etiquetas.
k) Estimación de los géneros a partir de los coeficientes de regresión obtenidos.
l) Evaluación del número de aciertos y de la tasa de aciertos en función de los AR.
m) Voting de las canciones del subconjunto de test para asignar la etiqueta.
n) Evaluación del número de aciertos en función del número de canciones, comparando etiquetas estimadas con reales.
o) Cálculo de la tasa de acierto en función del número de canciones.

4.4. Resultados de los experimentos

Tras la fase de experimentación es necesario discernir en qué horizontes temporales se consigue un mejor desempeño del algoritmo. A continuación se mostrarán gráficamente los resultados obtenidos.

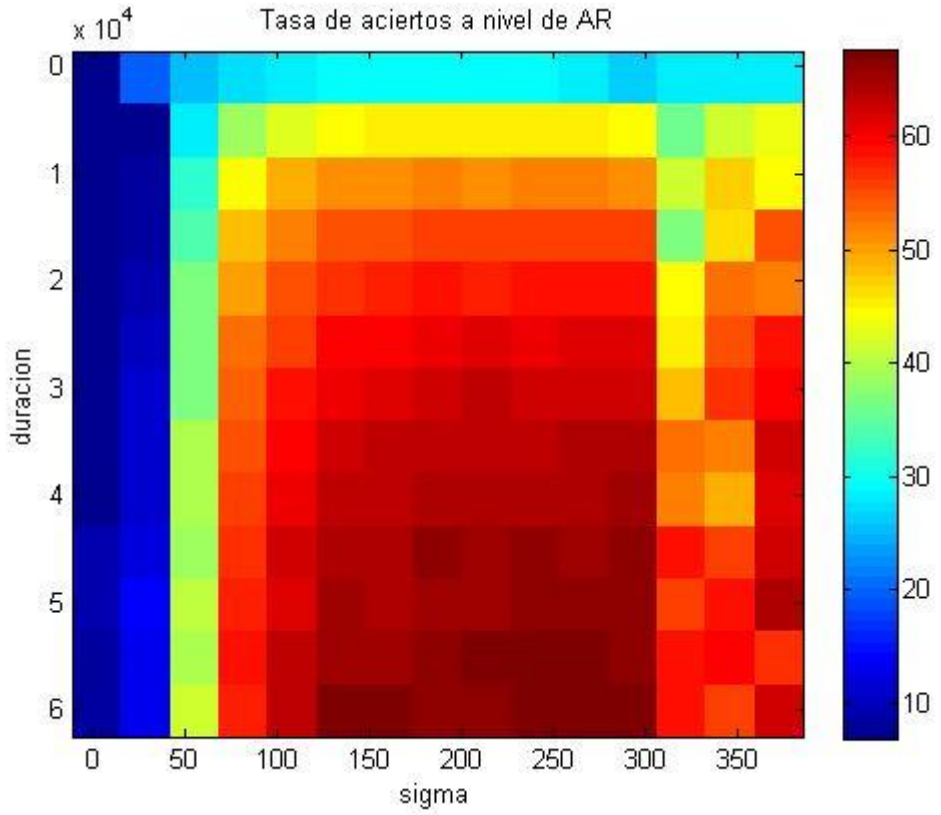
4.4.1. Resultados con la base de datos de artistas

4.4.1.1. Validación a nivel de AR

En la evaluación de aciertos a nivel de AR han de compararse los artistas estimados con la matriz de etiquetas de los datos de test. En ambas matrices el número de filas corresponde al número de coeficientes AR mientras que el número de columnas es el número de artistas, por tanto la clase a la que pertenezca vendrá definida por la posición que ocupe el máximo. En el caso de la matriz de artistas estimados el máximo será el resultado de la regresión lineal, mientras que en la matriz de etiquetas la clase viene marcada por las posiciones de los '1' dentro de una matriz inicialmente de ceros. Basta por tanto con verificar si las posiciones son idénticas para concluir si la predicción es un acierto o no. Dado que existen cinco iteraciones y en cada una de ellas los datos de test cambian, es necesario promediar los datos para reflejar correctamente el porcentaje de aciertos.

Las representaciones gráficas a nivel de AR indican la tasa de aciertos en función del tamaño de la ventana temporal y de la anchura del núcleo.

Figura 21. Tasa de aciertos a nivel de AR para la base de datos de artistas



Tasa de aciertos a nivel de AR para la base de datos de artistas, presentando la tasa de aciertos a nivel de AR vs. la duración y el valor de σ. En el eje de abscisas se representan los 15 valores de σ para los cuales se ha ejecutado el algoritmo. En el eje de ordenadas se representa el tamaño de la ventana temporal. Otra interpretación es considerar la gráfica como una matriz en la que se almacenan las tasas de acierto y cuyas filas representan la duración y las columnas los valores de σ.

En la Figura 21 se puede comprobar para qué posición dentro del vector de σ se alcanza el máximo número de aciertos a nivel de AR para cada duración. En este caso, los máximos se alcanzan en las siguientes posiciones:

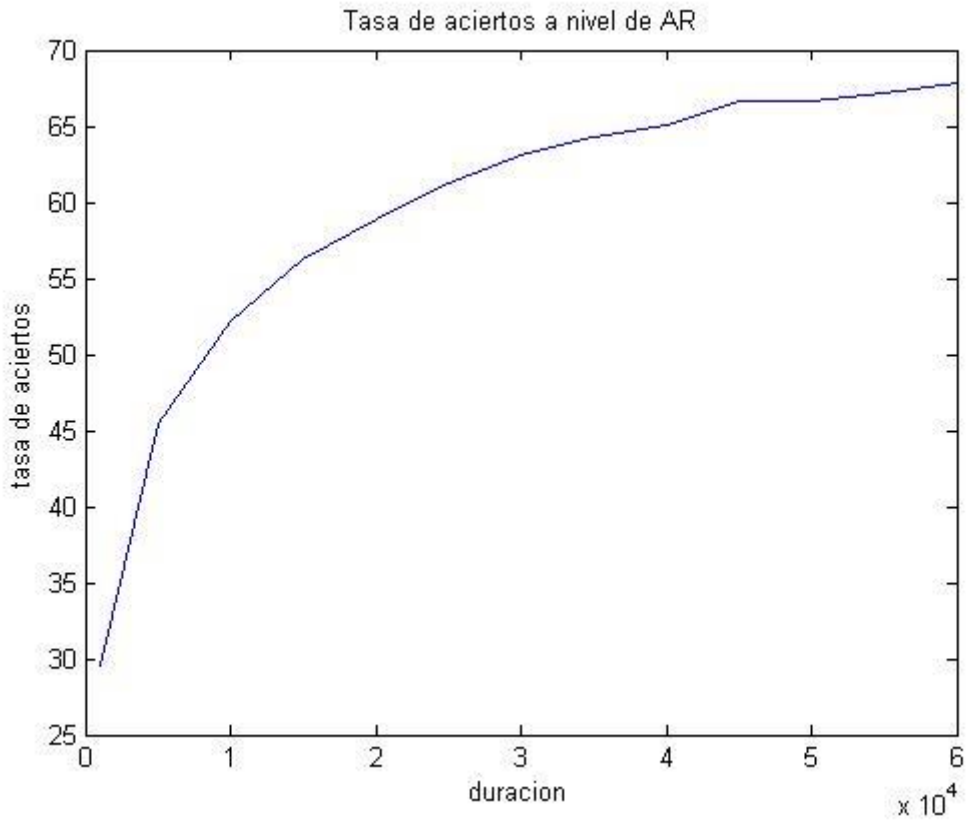
7	**8**	**10**	**11**	**12**	**11**	**9**	**12**	**12**	**12**	**10**	**9**	**7**

Los valores de σ en los que se produce mejor desempeño es en los comprendidos entre las posiciones 7 y 12, es decir, √(dim*4) y √(dim*128), que son valores relativamente centrados dentro de la escala con la que se han ejecutado los experimentos.

El valor de σ con el que el número de aciertos es mayor un mayor número de veces es $\sqrt{}$(dim*128). Este valor, relativamente elevado dentro del rango que se ha manejado, implica que aunque las muestras que se deben comparar no estén muy cercanas (en el sentido de similitud), una regresión lineal medianamente buena puede llevar a unas tasas de acierto alrededor del 60% en ventanas temporales iguales o superiores a los 10 segundos. Obviamente, cuanto mayor sea la ventana temporal mayor será el número de coeficientes, por tanto más finamente pueden extraerse las características. Esto queda corroborado con las tasas de acierto en la posición 12 de σ (valor $\sqrt{}$(dim*128)), notablemente superiores al 60%.

Como puede observarse, la validación es razonablemente estable, es decir, las máximas tasas de acierto se sitúan en una zona definida y relativamente delimitada en las posiciones centrales. Atendiendo al tamaño de la ventana temporal, tiempos de escucha inferiores a los 10 segundos suelen dar un porcentaje de acierto inferior al 50%. Esta relativa estabilidad queda también corroborada con la Figura 22. En dicha gráfica, para la σ óptima en cada momento, se identifica el máximo valor de acierto para cada duración.

Figura 22. Tasa de aciertos a nivel de AR para la base de datos de artistas



Tasa de aciertos a nivel de AR para la base de datos de artistas. Se representa el valor de la máxima tasa de acierto en función de la duración. Cada punto se extrae con la σ que se obtuvo por validación cruzada para cada tamaño de ventana para la integración temporal.

Los valores de la tasa de acierto máxima son:

29,5	**45,5**	**52,2**	**56,2**	**58,9**	**61,3**	**63,1**	**64,3**	**65,1**	**66,6**	**66,6**	**67,1**	**67,8**

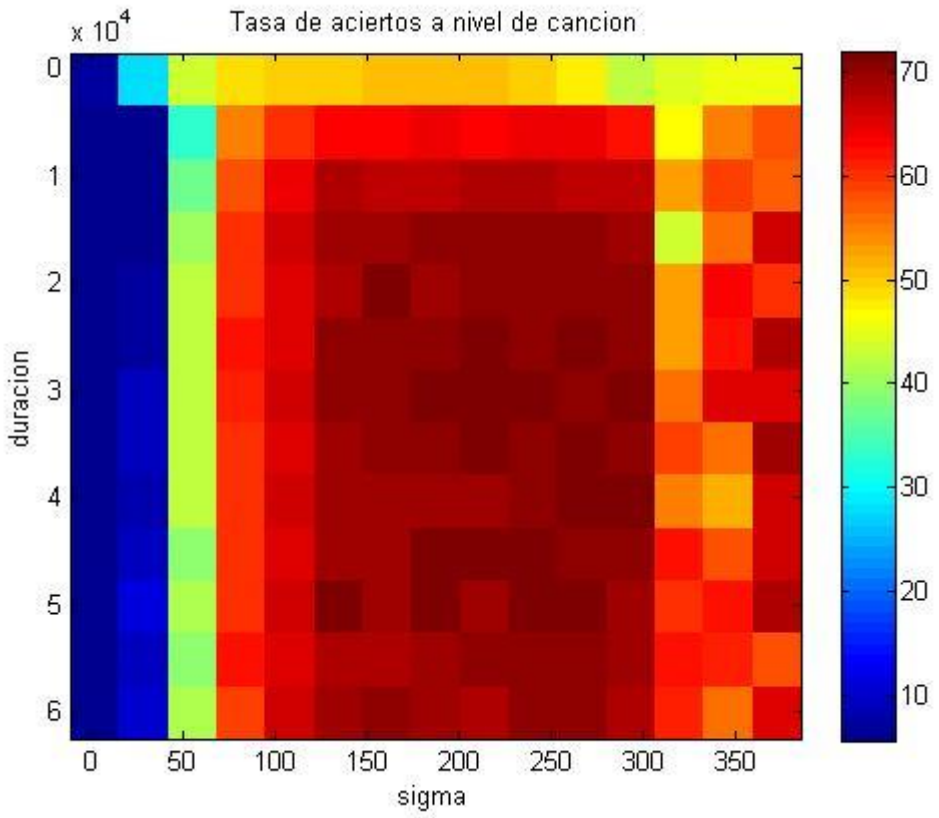
Cada uno de estos valores corresponde a la σ con la que se han alcanzado los valores máximos de tasa de acierto en cada punto. La tasa de aciertos es creciente a medida que aumenta la duración. Sin embargo, con duraciones menores la pendiente de la gráfica es mayor, y va disminuyendo a partir de los 10 segundos. Esto es porque en tamaños de ventana temporal menores un segundo adicional proporciona más información que un segundo añadido con una ventana temporal mayor.

4.4.1.2. Validación a nivel de canción

En el caso de la validación a nivel de canción, la comparación se lleva a cabo entre el vector con los resultados de la asignación de artista mediante voting para cada canción con el vector de etiquetas de los datos de test. Nuevamente los datos han de promediarse, esta vez en función del número de canciones existentes en dichos datos de test.

Las representaciones gráficas a nivel de canción indican la tasa de aciertos en función del tamaño de la ventana temporal.

Figura 23. Tasa de aciertos a nivel de canción para la base de datos de artistas



Tasa de aciertos a nivel de canción para la base de datos de artistas, presentando la tasa de aciertos a nivel de canción vs. la duración y el valor de σ. En el eje de abscisas se representan los 15 valores de σ para los cuales se ha ejecutado el algoritmo. En el eje de ordenadas se representa el tamaño de la ventana temporal. Otra interpretación es considerar la gráfica como una matriz en la que se almacenan las tasas de acierto y cuyas filas representan la duración y las columnas los valores de σ.

De nuevo puede observarse estabilidad en los datos, pues la zona de mayor tasa de aciertos vuelve a concentrarse en las posiciones centrales de σ. Esta vez las posiciones de σ en las que se dan los máximos son:

8	**8**	**9**	**11**	**7**	**9**	**9**	**9**	**11**	**8**	**6**	**9**	**10**

El mejor desempeño se obtiene en los valores comprendidos entre las posiciones 6 y 11, es decir, √(dim*2) y √(dim*64), que vuelven a ser valores relativamente centrados dentro de la escala con la que se han ejecutado los experimentos.

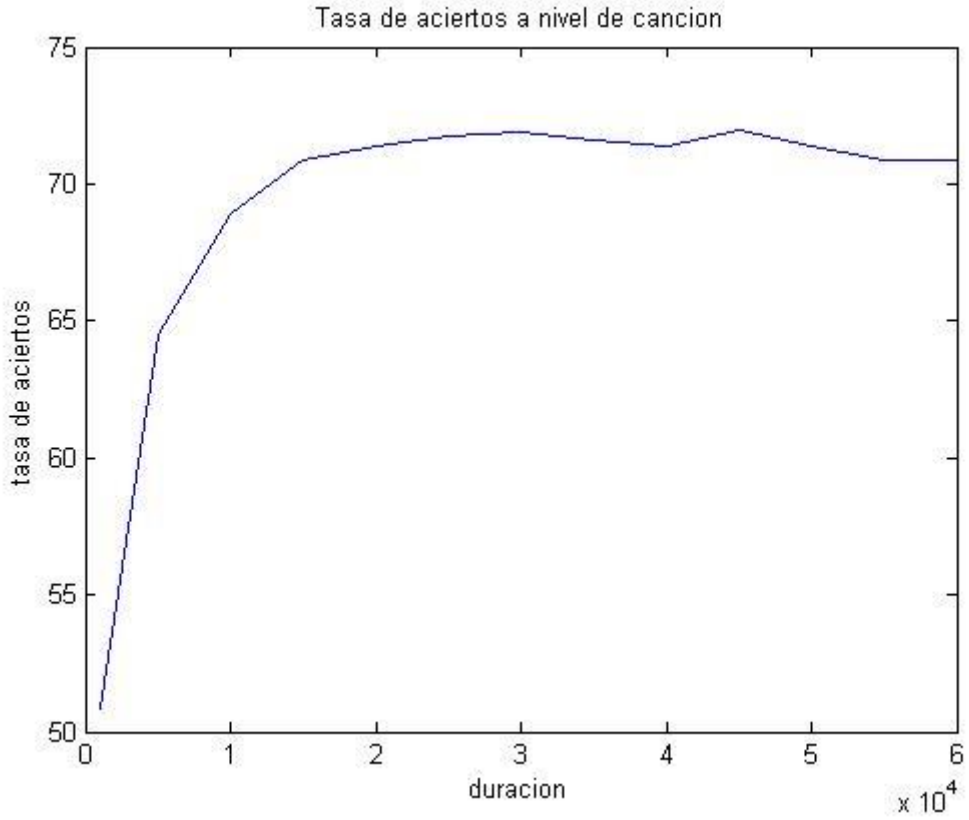
El valor de σ con el que el número de aciertos es mayor un mayor número de veces es √(dim*16). Este valor es sensiblemente menor que en el caso de la validación de AR, lo cual indica que para el reconocimiento de canciones las muestras entre las que se establece la comparación deben estar más cercanas entre sí para obtener mejores resultados.

La gráfica de la tasa de aciertos a nivel de canción en función de la duración se muestra a continuación. Los valores que alcanza son los siguientes:

50,9	**64,5**	**68,9**	**70,9**	**71,4**	**71,8**	**71,9**	**71,6**	**71,4**	**72**	**71,4**	**70,9**	**70,9**

Al igual que en el caso anterior, la tasa de aciertos también es creciente conforme aumenta la duración, pero sólo hacia los valores centrales del vector, a partir de ahí el crecimiento es más errático, produciéndose incluso momentos de estabilización y de ligero decrecimiento. Nótese también que para duraciones menores a los 10 segundos la pendiente es marcadamente más pronunciada que en el caso anterior. A la vista de los resultados se desprende que en la validación a nivel de canción se obtienen mejores resultados con ventanas temporales menores, si bien luego el crecimiento es más estable.

Figura 24. Tasa de aciertos a nivel de canción para la base de datos de artistas



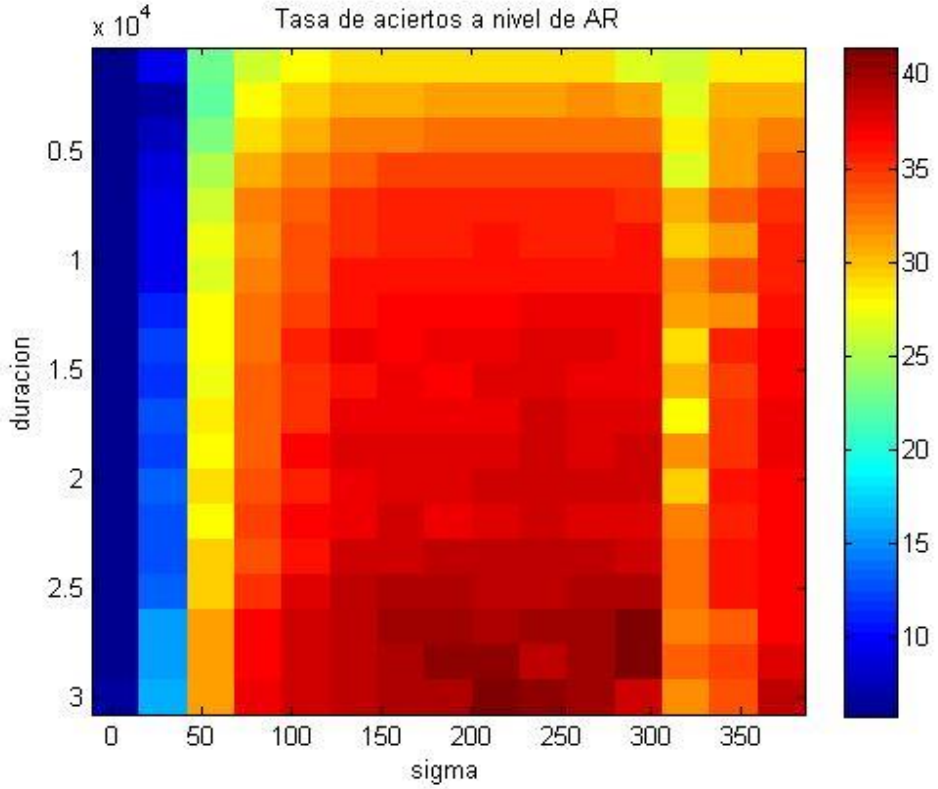
Tasa de aciertos a nivel de canción para la base de datos de artistas. Se representa el valor de la máxima tasa de acierto en función de la duración. Cada punto se extrae con la σ que se obtuvo por validación cruzada para cada tamaño de ventana para la integración temporal.

4.4.2. Resultados con la base de datos de género

4.4.2.1. Validación a nivel de AR

Esta vez en la evaluación de aciertos a nivel de AR han de compararse los géneros estimados con la matriz de etiquetas de los datos de test. El proceso es idéntico al seguido con la base de datos de artistas, en ambas matrices el número de filas corresponde al número de coeficientes AR mientras que el número de columnas es el número de géneros, por tanto la clase a la que pertenezca vendrá definida por la posición que ocupe el máximo. En el caso de la matriz de géneros estimados el máximo será el resultado de la regresión lineal, mientras que en la matriz de etiquetas la clase viene marcada por las posiciones de los '1' dentro de una matriz inicialmente de ceros. La verificación y el promedio de los datos también permanece igual.

Figura 25. Tasa de aciertos a nivel de AR para la base de datos de géneros



Tasa de aciertos a nivel de AR para la base de datos de géneros, presentando la tasa de aciertos a nivel de AR vs. la duración y el valor de σ. En el eje de abscisas se representan los 15 valores de σ para los cuales se ha ejecutado el algoritmo. En el eje de ordenadas se representa el tamaño de la ventana temporal. Otra interpretación es considerar la gráfica como una matriz en la que se almacenan las tasas de acierto y cuyas filas representan la duración y las columnas los valores de σ.

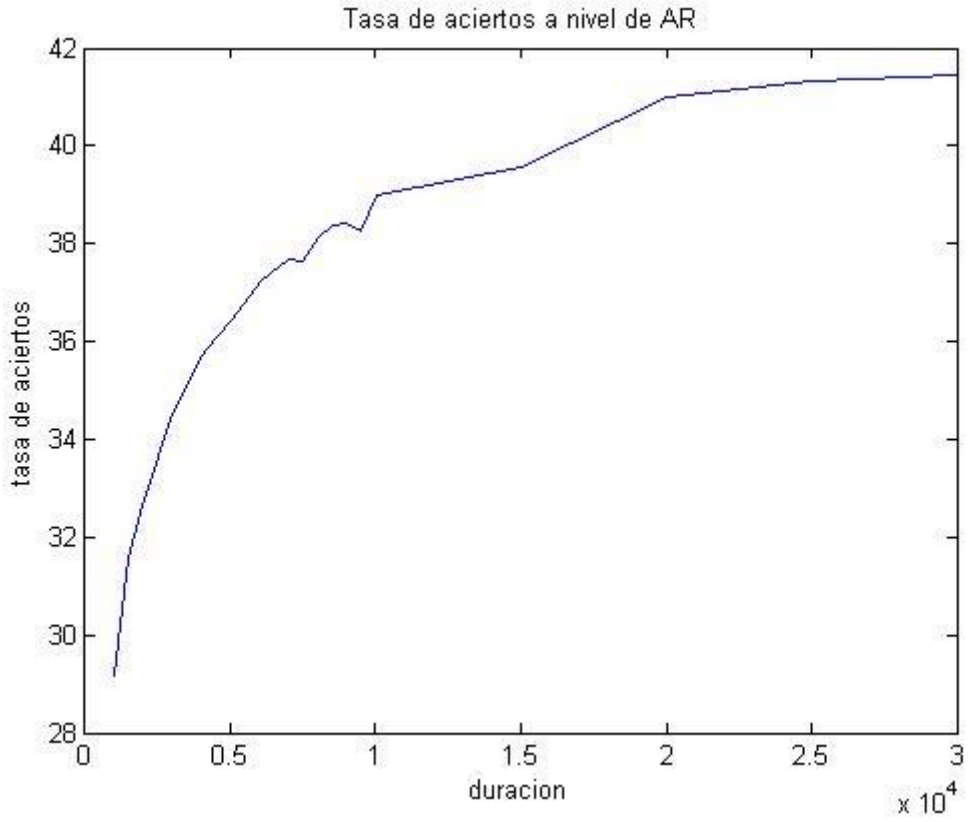
Los datos vuelven a concentrarse en las posiciones centrales de σ, entre la posición 7 y la 12, con unos valores de σ comprendidos entre $\sqrt{(dim*4)}$ y $\sqrt{(dim*128)}$, pero esta vez de forma menos delimitada y definida que en los experimentos con la base de datos de artistas. No obstante, sigue presentando una relativa estabilidad tanto en la propia selección de σ. Por otro lado, en la Figura 25 puede apreciarse también que la tasa de aciertos es menor con esta base de datos, mientras que con la base de datos de artistas, tanto en la validación por AR como por canción la tasa máxima de acierto se sitúa en aproximadamente un 70%, aquí encontramos una tasa aproximada del 40%. En este caso, las posiciones de σ en las que se sitúan las máximas tasas de acierto para cada duración son las siguientes:

10	**11**	**11**	**8**	**8**	**9**	**9**	**12**	**11**	**9**	**10**	**10**	**10**	**7**	**8**	**7**	**12**	**12**	**9**

Respecto a los valores de las tasas de acierto correspondientes a los máximos, puede observarse un crecimiento sostenido de ellos, manteniéndose la relativa estabilidad también en los resultados a nivel de AR. Se confirma también el valor sensiblemente menor que presentan con respecto al experimento anterior. Nótese también que, tal y como viene ocurriendo en los casos que se han presentado hasta ahora, la pendiente de la gráfica desde el inicio hasta aprox. los 7,5 segundos es más pronunciada.

29,2	**31,6**	**32,7**	**34,5**	**35,7**	**36,1**	**36,4**	**37,2**	**37,7**	**37,7**	**38,1**	**38,4**	**38,4**	**38,3**	**39**	**39,6**	**41**	**41,3**	**41,4**

Figura 26. Tasa de aciertos a nivel de AR para la base de datos de géneros

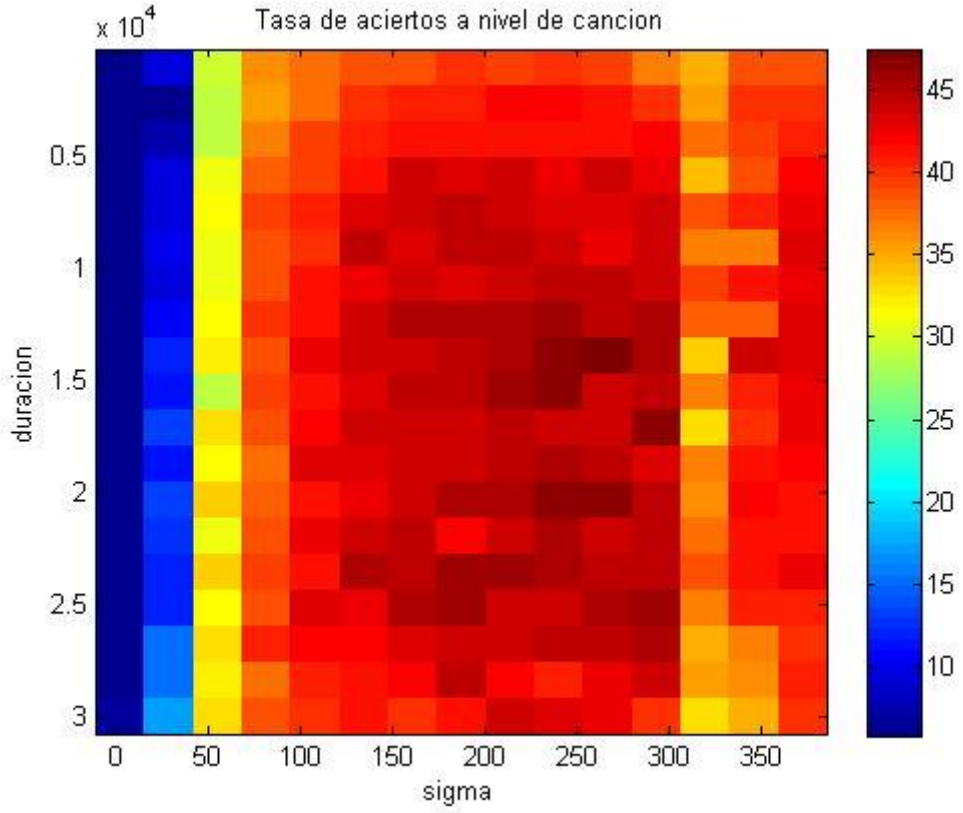


Tasa de aciertos a nivel de AR para la base de datos de géneros. Se representa el valor de la máxima tasa de acierto en función de la duración. Cada punto se extrae con la σ que se obtuvo por validación cruzada para cada tamaño de ventana para la integración temporal.

4.4.2.2. Validación a nivel de canción

En el caso de la validación a nivel de canción, la comparación se lleva a cabo entre el vector con los resultados de la asignación de género mediante voting para cada canción con el vector de etiquetas de los datos de test. Los datos también han de ser promediados en función del número de canciones existentes en dichos datos de test.

Figura 27. Tasa de aciertos a nivel de canción para la base de datos de géneros

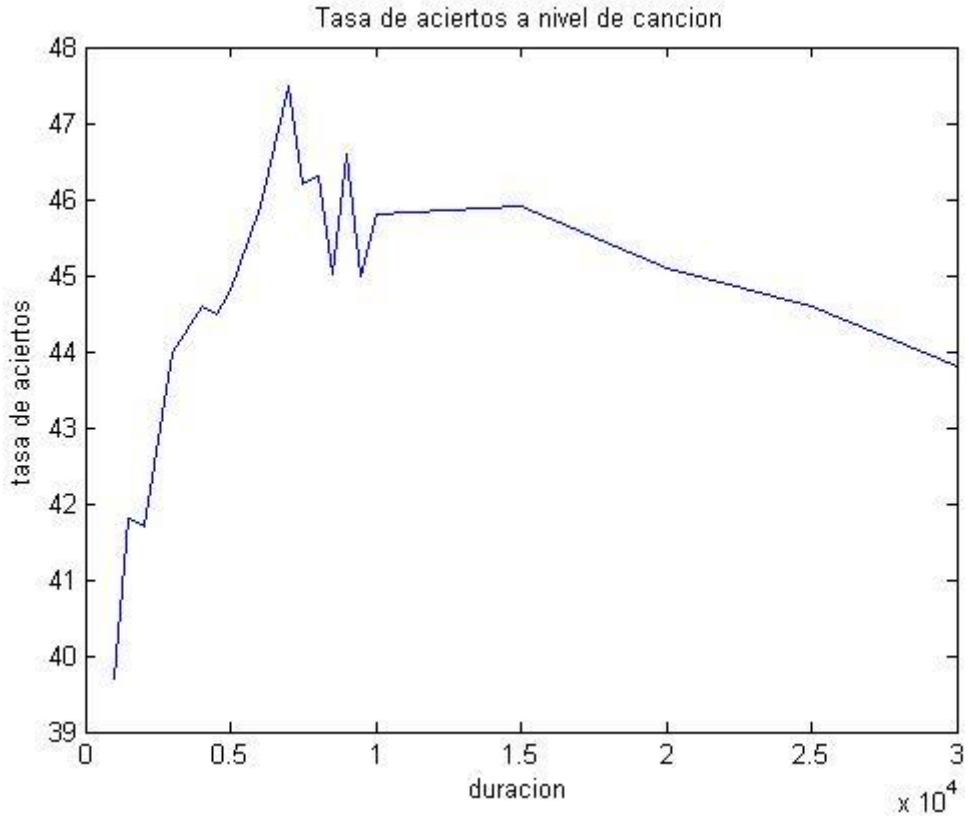


Tasa de aciertos a nivel de canción para la base de datos de géneros, presentando la tasa de aciertos a nivel de canción vs. la duración y el valor de sigma. En el eje de abscisas se representan los 15 valores de σ para los cuales se ha ejecutado el algoritmo. En el eje de ordenadas se representa el tamaño de la ventana temporal. Otra interpretación es considerar la gráfica como una matriz en la que se almacenan las tasas de acierto y cuyas filas representan la duración y las columnas los valores de σ.

Este último experimento también presenta la mayor tasa de acierto en las posiciones centrales de σ, sin embargo continúa mostrando menor uniformidad en la distribución de las tasas máximas que los experimentos realizados con la base de datos de artistas. Los valores de las posiciones de σ en este experimento están comprendidas entre 8 y 12, cuyos valores de σ son $\sqrt{(dim*2)}$ y $\sqrt{(dim*128)}$:

10	**9**	**12**	**11**	**8**	**8**	**10**	**10**	**11**	**10**	**12**	**10**	**11**	**10**	**9**	**12**	**12**	**8**	**9**

Figura 28. Tasa de aciertos a nivel de canción para la base de datos de géneros



Tasa de aciertos a nivel de canción para la base de datos de géneros. Se representa el valor de la máxima tasa de acierto en función de la duración. Cada punto se extrae con la σ que se obtuvo por validación cruzada para cada tamaño de ventana para la integración temporal.

Respecto a la validación representada en la Figura 28, puede verse que la tasa de reconocimiento es alta para ventanas de integración en torno a los 10 segundos para luego decrecer. Se repite también el patrón de pendiente pronunciada en torno a los 7,5 segundos. De los experimentos realizados, este es el que presenta menor uniformidad en los datos. En esta ocasión las máximas tasas de acierto en función de la duración son las siguientes:

39,7	41,8	41,7	44	44,6	44,5	44,8	45,9	47,5	46,2	46,3	45	46,6	45	45,8	45,9	45,1	44,6	43,8

Es de destacar que, si bien se repite el patrón de pendiente pronunciada en ventanas temporales menores, la máxima tasa de aciertos es apenas 8 puntos porcentuales mayor que con la ventana temporal de menor tamaño.

CAPÍTULO 5. Conclusiones y Líneas Futuras

En este trabajo hemos abordado el problema de la clasificación automática de canciones en categorías de alto nivel, como son el género y el artista, utilizando para ello únicamente la propia señal de audio. Estos sistemas son de gran interés en tareas de organización automática y recomendación, y su estudio ha experimentado un gran avance como consecuencia de la distribución digital y el abaratamiento de los dispositivos de almacenamiento y reproducción.

La mayoría de los sistemas que abordan estas tareas se basan en las siguientes etapas, que han sido debidamente explicadas a lo largo de la memoria:

- Extracción de descriptores de bajo nivel (típicamente MFCCs) en ventanas de corta duración (20-40 ms) en los que se asume que la señal es estacionaria.
- Integración temporal de características: se trata de incorporar la información de todos los MFCCs extraídos en una ventana de mayor duración temporal en un único vector, de manera tal que se conserve e incluso se haga más evidente las características más relevantes para la tarea de clasificación abordada.
- Clasificación de los vectores de mayor escala temporal en base a las categorías predefinidas. En este proyecto se han descrito y utilizado para esta fase clasificadores basados en métodos núcleo.

El presente trabajo ha estudiado la influencia de la etapa de integración temporal en las prestaciones globales del sistema, concretamente analizando la evolución de la tasa de error al modificar el tamaño de la ventana de mayor duración. En trabajos previos, se han utilizado tamaños de ventana que varían desde 1 a 30 sg, argumentando que únicamente a esta escala temporal se puede extraer la información de interés. En este trabajo hemos llevado a cabo un rastreo más cuidadoso de dicha duración para dos tareas diferentes de clasificación: género musical y artista.

Las principales conclusiones que hemos obtenido de nuestro estudio son:

- El valor óptimo de la anchura del núcleo gaussiano parece no ser muy dependiente del tamaño de ventana escogido.
- Incrementar el tamaño de ventana ha resultado beneficioso, para ambas bases de datos, respecto de la tasa de acierto obtenida sobre los vectores AR resultantes de la integración temporal. En otras palabras, mientras mayor sea la escala temporal utilizada para la extracción de vectores AR, mayor precisión tendremos al asignar dichos vectores a su género musical o artista correspondiente.
- Como contrapartida de lo anterior, cuando medimos la tasa de acierto sobre canciones, tomando la decisión con voto por mayoría sobre las decisiones correspondientes a todos los vectores AR, parece existir un tamaño de ventana óptimo, a partir del cual la precisión del clasificador decrece (aunque no muy significativamente). Dicho deterioro en prestaciones puede deberse únicamente a que el número de coeficientes AR disponibles para la clasificación de cada canción disminuye al aumentar el tamaño de la ventana de integración temporal.

- Las conclusiones parecen bastante independientes de la base de datos utilizada, si bien el deterioro de prestaciones sufrido al incrementar el tamaño de ventana por encima de la duración óptima parece más importante en la tarea de clasificación de género musical.

A la vista de los resultados, podemos concluir que las tareas de validación cruzada resultan críticas a la hora de diseñar un sistema de clasificación automática, no únicamente en lo relativo a los parámetros del clasificador, sino también en cuanto al tamaño de ventana utilizado en la fase de integración temporal. Otros parámetros como por ejemplo el número de coeficientes MFCC parecen tener una importancia menor, atendiendo a los resultados disponibles en la literatura.

Como líneas futuras de trabajo se sugiere extender el trabajo de simulación a otras bases de datos, otros descriptores de bajo nivel, e incluso otras características de integración temporal, como pueden ser los periodogramas. El objetivo de dicho trabajo sería comprobar si el tamaño óptimo de ventana depende únicamente de la tarea abordada, o por el contrario el tipo de descriptores y características utilizados por el sistema también influye en dicho valor.

Por último, una línea de trabajo que también se considera de interés es utilizar el tamaño de la ventana de integración temporal como una forma de incorporar diversidad y en última instancia mejorar las prestaciones del sistema. Así, por ejemplo, si distintos géneros musicales se viesen mejor representados a distinta escala temporal, un sistema "multiescala" podría aprovechar el uso de ventanas de distinto tamaño para ofrecer una mayor precisión global.

APÉNDICE 1. Bibliografía Ordenada

[Ahrendt et al., 2004]

Ahrendt, P., Meng, A., Larsen, J., "Decision time horizon for music genre classification using short-time features" EUSIPCO, 2004, pp. 1293-1296.

[Arenas-García et al., 2006]

Arenas-García, J., Petersen, K.B., Hansen, L.K. "Sparse Kernel Orthonormalized PLS for Feature Extraction in Large Data Sets", Informatics and Mathematical Modelling, Technical University of Denmark, December 2006.

[Arenas-García et al., 2006] (Incluir en capítulos 3 y 4)

Arenas-García, J., Meng, A., Petersen, K. B., y Hansen, L. K. "Multivariate Analysis and Kernel Methods for Music Data Analysis", Advances in models for Acoustic Processing workshop, NIPS'06, Whistler, Canada, 2006.

[Bishop, 1995]

Bishop, C.M., - 1995. "Neural Networks for Pattern Recognition". Oxford University Press.

[Bousquet y Pérez-Cruz, 2003]

Bousquet, O., Pérez-Cruz, F., "Kernel Methods and Their Application to Signal Processing". Max Planck Institute, pdf 2018. January 2003.

[Detyniecki et al., 2005]

Detyniecki, M., Jose, J.M., Nümberger, A., van Rijsbergen, C.J., "Adaptative Multimedia Retrieval: User, Context and Feedback". Third International Workshop, AMR 2005, Glasgow, UK, 2005. Ed. Springer, 2005.

[Ellis, 2007]

Ellis, Daniel P. W., "Classifying Music Audio with Timbral and Chroma Features". Proc. Int. Conf. on Music Information Retrieval ISMIR-07, Vienna, Austria, Sep. 2007.

[Gook y Sikora, 2004]

Gook, K. H., Sikora, T., "Audio Spectrum projection based on several basis decomposition algorithms applied to general sound recognition and audio segmentation" EUSIPCO, 2004, pp. 1047-1050.

[Hayes, 1996]

Hayes, M. H., "Statistical Digital Signal Processing and Modeling", John Wiley & Sons, New York, 1996.

[Haykin, 1999]

Haykin, S., - 1999. "Neural Networks. A Comprehensive Foundation". Second Edition. International Edition. Prentice Hall.

[Jurafsky y Martin, 2008]

Jurafsky, D., Martin, J.H. "Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition". Prentice Hall, 2008.

[Kim et al., 2005]

Kim, H.-K., Moreau, N. y Sikora, T. - 2005. MPEG-7 Audio and Beyond. Audio Context Indexing and Retrieval. John Wiley & Sons, Ltd. Páginas 2:10.

[Logan, 2000]

Logan, B. "Mel Frequency Cepstral Coefficients for Music Modelling", en Proc. Int. Symp. Music Inf. Retrieval, Plymouth, MA, Octubre 2000.

[Loughran et al., 2008]

Loughran, R., Walker, J., O'Neill, M. y O'Farrell, M. "The Use of Mel-frequency Cepstral Coefficients in Musical Instrument Identification", *Proc. ICMC 2008*, 24-29 Ago. 2008, Belfast, Northern Ireland

[Meng et al., 2005]

Meng, A., Ahrendt, P., Larsen, J. "Improving Music Genre Classification by Short Time Feature Integration", International Conference on Acoustics, Speech and Signal Processing, Volume 5, pp. 497-500

[Meng y Shawe-Taylor, 2005]

Meng, A., Shawe-Taylor, J., "An Investigation of Feature Models for Music Genre Classification using the Support Vector Classifier"

[Meng, 2006]

Meng, A. "Temporal Feature Integration for Music Organisation" Ph. D. thesis – Kongens Lyngby 2006. IMM-PHD-2006-165 ISSN 0909-3192

[Meng et al., 2007]

Meng, A. Ahrendt, P. Larsen, J. Hansen, L.K., "Temporal Feature Integration for Music Genre Classification", IEEE Transactions on Audio, Speech and Language Processing, Volume 15, Issue 5, pp. 1654-1664 July 2007

[Nielsen et al., 2007]

Nielsen, A.B., Sigurdsson, S., Hansen, L.K., Arenas-García, J., "On the Relevance of Spectral Features for Instrument Classification**",** IEEE International Conference on Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. Volume 2, pp. II-485-II-488. April 2007.

[Oppenheim y Schafer, 1989]

Oppenheim, A.V., Schafer, R. W. "Discrete-Time Signal Processing", Englewood Cliffs, NJ:. Prentice-Hall, 1989.

[Peña Sánchez de Rivera, 1992]

Peña Sánchez de Rivera, D., "Modelos y Métodos. Modelos Lineales y Series Temporales". Alianza Universidad, Madrid, 1992

[Rosipal y Trejo, 2001]

Rosipal, R., Trejo, L. J., "Kernel Partial Least Squares regresión in Reproducing Kernel Hilbert Space" Journal of Machine Learning Research, 2001: 97-123.

[Schölkopf et al., 1999]

Schölkopf, B., Burges, C. J. C., Smola, A. J. "Advances in Kernel Methods--Support Vector Learning". MIT Press, Cambridge, MA, 1999.

[Schölkopf y Smola, 2002]

Schölkopf, B., Smola, A. J. "Learning with Kernels". MIT Press, Cambridge, MA, London, England 2002.

[Shawe-Taylor y Cristianini, 2004]

Shawe-Taylor, J. y Cristianini, N. "Kernel Methods for Pattern Analysis" Cambridge University Press, 2004.

[Sigurdsson et al., 2006]

Sigurdsson, S, Petersen, K.B. y Lehn-Schioler, T. "Mel Frequency Cepstral Coefficient: An Evaluation of Robustness of mp3 Encoded Music", en ISMIR 2006, Victoria, Canadá, 2006.

[Vapnik, 2000]

Vapnik, V.N. "The nature of statistical learning theory". Springer, 2000.

APÉNDICE 2. Agradecimientos

En primer lugar, me gustaría agradecerle a mi tutor, Jerónimo Arenas García, todo lo que me ha aportado desde el día en que lo conocí. Si bien la inteligencia es una cualidad, y como cualidad se nace con ella y sólo de la propia persona depende cultivarla o no, la brillantez y la generosidad de compartir lo aprendido es lo verdaderamente meritorio y digno de agradecer. Muchas gracias por tu capacidad de hacer un poco más fácil lo difícil, por tu paciencia, por tu dedicación y por tu inestimable ayuda no sólo en este Proyecto, sino también en todas las tutorías en las que me has recibido siempre con una sonrisa. Muchas gracias por todo Jero.

También quisiera agradecer su ayuda a mi amiga Sara Gonell, por guiarme en mi camino de vuelta de la desmemoria.

Por último, y no menos importante, quiero dedicar este Proyecto a mis padres, por su infinito amor y su ejemplo constante.